Object-Centric Video Representation for Long-Term Action Anticipation

Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, Chen Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6751-6761

This paper focuses on building object-centric representations for long-term action anticipation in videos. Our key motivation is that objects provide important cues to recognize and predict human-object interactions, especially when the predictions are longer term, as an observed "background" object could be used by the human actor in the future. We observe that existing object-based video recognition frameworks either assume the existence of in-domain supervised object detectors or follow a fully weakly-supervised pipeline to infer object locations from action labels. We propose to build object-centric video representations by leveraging visual-language pretrained models. This is achieved by "object prompts", an approach to extract task-specific object-centric representations from general-purpose pretrained models without finetuning. To recognize and predict human-object interactions, we use a Transformer-based neural architecture which allows the "retrieval" of relevant objects for action anticipation at various time scales. We conduct extensive evaluations on the Ego4D, 50Salads, and EGTEA Gaze+ benchmarks. Both quantitative and qualitative results confirm the effectiveness of our proposed method.

****

CLRerNet: Improving Confidence of Lane Detection With LaneIoU

Hiroto Honda, Yusuke Uchida; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1176-1185

Lane marker detection is a crucial component of the autonomous driving and driver assistance systems. Modern deep lane detection methods with anchor-based lane representation exhibit excellent performance on lane detection benchmarks. Through preliminary oracle experiments, we firstly disentangle the lane representation components to determine the direction of our approach. We show that correct lane positions are already among the predictions of an existing anchor-based detector, and the confidence scores that accurately represent intersection-over-union (IoU) with ground truths are the most beneficial. Based on the finding, we propose LaneIoU that better correlates with the metric, by taking the local lane angles into consideration. We develop a novel detector coined CLRerNet featuring LaneIoU for the target assignment cost and loss functions aiming at the improved quality of confidence scores. Through careful and fair benchmark including cross validation, we demonstrate that CLRerNet outperforms the state-of-the-art by a large margin - enjoying F1 score of 81.43% compared with 80.47% of the existing method on CULane, and 86.47% compared with 86.10% on CurveLanes.

****

Training Ensembles With Inliers and Outliers for Semi-Supervised Active Learning

Vladan Stojni■, Zakaria Laskar, Giorgos Tolias; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 260-269

Deep active learning in the presence of outlier examples poses a realistic yet challenging scenario. Acquiring unlabeled data for annotation requires a delicate balance between avoiding outliers to conserve the annotation budget and prioritizing useful inlier examples for effective training. In this work, we present an approach that leverages three highly synergistic components, which are identified as key ingredients: joint classifier training with inliers and outliers, semi-supervised learning through pseudo-labeling, and model ensembling. Our work demonstrates that ensembling significantly enhances the accuracy of pseudo-labeling and improves the quality of data acquisition. By enabling semi-supervision through the joint training process, where outliers are properly handled, we observe a substantial boost in classifier accuracy through the use of all available unlabeled examples. Notably, we reveal that the integration of joint training renders explicit outlier detection unnecessary; a conventional component for acquisition in prior work. The three key components align seamlessly with numerous existing approaches. Through empirical evaluations, we showcase that their combined use leads to a performance increase. Remarkably, despite its simplicity, our proposed approach outperforms all other methods in terms of performance. Code: https:

//github.com/vladan-stojnic/active-outliers

****

Robust Source-Free Domain Adaptation for Fundus Image Segmentation

Lingrui Li, Yanfeng Zhou, Ge Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7840-7849

Unsupervised Domain Adaptation (UDA) is a learning technique that transfers knowledge learned in the source domain from labelled training data to the target domain with only unlabelled data. It is of significant importance to medical image segmentation because of the usual lack of labelled training data. Although extensive efforts have been made to optimize UDA techniques to improve the accuracy of segmentation models in the target domain, few studies have addressed the robustness of these models under UDA. In this study, we propose a two-stage training strategy for robust domain adaptation. In the source training stage, we utilize adversarial sample augmentation to enhance the robustness and generalization capability of the source model. And in the target training stage, we propose a novel robust pseudo-label and pseudo-boundary (PLPB) method, which effectively utilizes unlabeled target data to generate pseudo labels and pseudo boundaries that enable model self-adaptation without requiring source data. Extensive experimental results on cross-domain fundus image segmentation confirm the effectiveness and versatility of our method. Source code of this study is openly accessible at https://github.com/LinGrayy/PLPB.

****

Controlling Rate, Distortion, and Realism: Towards a Single Comprehensive Neural Image Compression Model

Shoma Iwai, Tomo Miyazaki, Shinichiro Omachi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2900-2909

In recent years, neural network-driven image compression (NIC) has gained significant attention. Some works adopt deep generative models such as GANs and diffusion models to enhance perceptual quality (realism). A critical obstacle of these generative NIC methods is that each model is optimized for a single bit rate. Consequently, multiple models are required to compress images to different bit rates, which is impractical for real-world applications. To tackle this issue, we propose a variable-rate generative NIC model. Specifically, we explore several discriminator designs tailored for the variable-rate approach and introduce a novel adversarial loss. Moreover, by incorporating the newly proposed multi-realism technique, our method allows the users to adjust the bit rate, distortion, and realism with a single model, achieving ultra-controllability. Unlike existing variable-rate generative NIC models, our method matches or surpasses the performance of state-of-the-art single-rate generative NIC models while covering a wide range of bit rates using just one model.

****

MetaVers: Meta-Learned Versatile Representations for Personalized Federated Learning

Jin Hyuk Lim, SeungBum Ha, Sung Whan Yoon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2587-2596

One of the daunting challenges in federated learning (FL) is the heterogeneity across clients that hinders the successful federation of a global model. When the heterogeneity becomes worse, personalized federated learning (PFL) pursues to detour the hardship of capturing the commonality across clients by allowing the personalization of models built upon the federation. In the scope of PFL for visual models, on the contrary, the recent effort for aggregating an effective global representation rather than chasing further personalization draws great attention. Along the same lines, we aim to train a large-margin global representation with a strong generalization across clients by adopting the meta-learning framework and margin-based loss, which are widely accepted to be effective in handling multiple visual tasks. Our method called MetVers achieves state-of-the-art accuracies for the PFL benchmarks with the CIFAR-10, CIFAR-100, and CINIC-10 datasets while showing robustness against data reconstruction attacks. Noteworthy, the versatile representation of MetaVers exhibits a strong generalization when tested on new clients with novel classes.

****

Improving Open-Set Semi-Supervised Learning With Self-Supervision

Erik Wallin, Lennart Svensson, Fredrik Kahl, Lars Hammarstrand; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2356-2365

Open-set semi-supervised learning (OSSL) embodies a practical scenario within semi-supervised learning, wherein the unlabeled training set encompasses classes absent from the labeled set. Many existing OSSL methods assume that these out-of-distribution data are harmful and put effort into excluding data belonging to unknown classes from the training objective. In contrast, we propose an OSSL framework that facilitates learning from all unlabeled data through self-supervision. Additionally, we utilize an energy-based score to accurately recognize data belonging to the known classes, making our method well-suited for handling uncurated data in deployment. We show through extensive experimental evaluations that our method yields state-of-the-art results on many of the evaluated benchmark problems in terms of closed-set accuracy and open-set recognition when compared with existing methods for OSSL. Our code is available at https://github.com/walline/ssl-tf2-sefoss.

****

FOSSIL: Free Open-Vocabulary Semantic Segmentation Through Synthetic References Retrieval

Luca Barsellotti, Roberto Amoroso, Lorenzo Baraldi, Rita Cucchiara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1464-1473

Unsupervised Open-Vocabulary Semantic Segmentation aims to segment an image into regions referring to an arbitrary set of concepts described by text, without relying on dense annotations that are available only for a subset of the categories. Previous works relied on inducing pixel-level alignment in a multi-modal space through contrastive training over vast corpora of image-caption pairs. However, representing a semantic category solely through its textual embedding is insufficient to encompass the wide-ranging variability in the visual appearances of the images associated with that category. In this paper, we propose FOSSIL, a pipeline that enables a self-supervised backbone to perform open-vocabulary segmentation relying only on the visual modality. In particular, we decouple the task into two components: (1) we leverage text-conditioned diffusion models to generate a large collection of visual embeddings, starting from a set of captions. These can be retrieved at inference time to obtain a support set of references for the set of textual concepts. Further, (2) we exploit self-supervised dense features to partition the image into semantically coherent regions. We demonstrate that our approach provides strong performance on different semantic segmentation datasets, without requiring any additional training.

****

Activity-Based Early Autism Diagnosis Using a Multi-Dataset Supervised Contrastive Learning Approach

Asha Rani, Yashaswi Verma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7788-7797

Autism Spectrum Disorder (ASD) is a neurological disorder. Its primary symptoms include difficulty in verbal/non-verbal communication and rigid/repetitive behavior. Traditional methods of autism diagnosis require multiple visits to a human specialist. However, this process is generally time-consuming and may result in a delayed (early) intervention. In this paper, we present a data-driven approach to automate autism diagnosis using video clips of subjects performing simple activities recorded in a weakly constrained environment. This task is particularly challenging since the available training data is small, videos from the two categories ("ASD" and "Control") are generally perceptually indistinguishable, and there is no clear understanding of what features would be beneficial in this task. To address these, we present a novel multi-dataset supervised contrastive learning technique to learn discriminative features simultaneously from multiple video datasets with significantly diverse distributions. Extensive empirical analyses demonstrate the promise of our approach compared to competing techniques on this challenging task.

****

Label Shift Estimation for Class-Imbalance Problem: A Bayesian Approach

Changkun Ye, Russell Tsuchida, Lars Petersson, Nick Barnes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1073-1082

As a type of distribution shift, label shift occurs when the source and target domains have different label distributions P(Y) but identical conditional distributions of data given labels P(X | Y). Under a Bayesian framework, we propose a novel Maximum A Posteriori (MAP) model and a novel posterior sampling model for the label shift problem. We prove the MAP objective admits a unique optimum and derive an EM algorithm that converges to the global optimum. We propose a novel Adaptive Prior Learning (APL) model to adaptively select prior parameters given data. We use the Markov Chain Monte Carlo (MCMC) method in our posterior sampling model to estimate and correct for label shift. Our methods can effectively resolve class imbalance problems on large-scale datasets without fine-tuning the classifier. Experiments show that our model outperforms existing methods on a variety of label shift settings. Our code is available at https://github.com/ChangkunYe/MAPLS/

****

SeaTurtleID2022: A Long-Span Dataset for Reliable Sea Turtle Re-Identification

Lukáš Adam, Vojt■ch ■ermák, Kostas Papafitsoros, Lukas Picek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7146-7156

This paper introduces the first public large-scale, long-span dataset with sea turtle photographs captured in the wild - SeaTurtleID2022. The dataset contains 8729 photographs of 438 unique individuals collected within 13 years, making it the longest-spanned dataset for animal re-identification. Each photograph includes various annotations, e.g., identity, encounter timestamp, and body parts segmentation masks. Instead of a standard "random" split, the dataset allows for two realistic and ecologically motivated splits: (i) time-aware: a closed-set with training, validation, and test data from different days/years, and (ii) open-set: with new unknown individuals in test and validation sets. We show that time-aware splits are essential for benchmarking methods for re-identification, as random splits lead to performance overestimation. Furthermore, a baseline instance segmentation and re-identification performance over various body parts is provided. At last, an end-to-end system for sea turtle re-identification is proposed and evaluated. The proposed system based on Hybrid Task Cascade for head instance segmentation and ArcFace-trained feature-extractor achieved an accuracy of 86.8%.

****

Self-Supervised Edge Detection Reconstruction for Topology-Informed 3D Axon Segmentation and Centerline Detection

Alec S. Xu, Nina I. Shamsi, Lars A. Gjesteby, Laura J. Brattain; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7831-7839

Many machine learning-based axon tracing methods rely on image datasets with segmentation labels. This requires manual annotation from domain experts, which is labor-intensive and not practical for large-scale brain mapping on hemisphere or whole brain tissue at cellular or sub-cellular resolution. Additionally, preserving axon structure topology is crucial to understanding neural connections and brain function. Self-supervised learning (SSL) is a machine learning framework that allows models to learn an auxiliary task on unannotated data to aid performance on a supervised target task. In this work, we propose a novel SSL auxiliary task of reconstructing an edge detector for the target task of topology-oriented axon segmentation and centerline detection. We pretrained 3D U-Nets on three different SSL tasks using a mouse brain dataset: our proposed task, predicting the order of permuted slices, and playing a Rubik's cube. We then evaluated these U-Nets and a baseline model on a different mouse brain dataset. Across all experiments, the U-Net pretrained on our proposed task improved the baseline's segmentation, topology-preservation, and centerline detection by up to 5.03%, 4.65%, an d 5.41%, respectively. In contrast, there was no consistent improvement over the baseline observed with the slice-permutation and Rubik's cube pretrained U-Nets.

****

Bi-Directional Training for Composed Image Retrieval via Text Prompt Learning

Zheyuan Liu, Weixuan Sun, Yicong Hong, Damien Teney, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5753-5762

Composed image retrieval searches for a target image based on a multi-modal user query comprised of a reference image and modification text describing the desired changes. Existing approaches to solving this challenging task learn a mapping from the (reference image, modification text)-pair to an image embedding that is then matched against a large image corpus. One area that has not yet been explored is the reverse direction, which asks the question, what reference image when modified as described by the text would produce the given target image? In this work we propose a bi-directional training scheme that leverages such reversed queries and can be applied to existing composed image retrieval architectures with minimum changes, which improves the performance of the model. To encode the bi-directional query we prepend a learnable token to the modification text that designates the direction of the query and then finetune the parameters of the text embedding module. We make no other changes to the network architecture. Experiments on two standard datasets show that our novel approach achieves improved performance over a baseline BLIP-based model that itself already achieves competitive performance. Our code is released at https://github.com/Cuberick-Orion/Bi-Blip4CIR.

****

iBARLE: imBalance-Aware Room Layout Estimation

Taotao Jing, Lichen Wang, Naji Khosravan, Zhiqiang Wan, Zachary Bessinger, Zhengming Ding, Sing Bing Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 914-924

Room layout estimation predicts layouts from a single panorama. It requires datasets with large-scale and diverse room shapes to well train the models. However, there are significant imbalances in real-world datasets including the dimensions of layout complexity, camera locations, and variation in scene appearance. These issues considerably influence the model training performance. In this work, we propose imBalance-Aware Room Layout Estimation (iBARLE) framework to address these issues. iBARLE consists of: (1) Appearance Variation Generation (AVG) module, which promotes visual appearance domain generalization, (2) Complex Structure Mix-up (CSMix) module, which enhances generalizability w.r.t. room structure, and (3) a gradient-based layout objective function, which allows more effective accounting for occlusions in complex layouts. All modules are jointly trained and help each other to achieve the best performance. Experiments and ablation studies based on ZInD dataset illustrate that iBARLE has state-of-the-art performance compared with other layout estimation baselines.

****

FarSight: A Physics-Driven Whole-Body Biometric System at Large Distance and Altitude

Feng Liu, Ryan Ashbaugh, Nicholas Chimitt, Najmul Hassan, Ali Hassani, Ajay Jaiswal, Minchul Kim, Zhiyuan Mao, Christopher Perry, Zhiyuan Ren, Yiyang Su, Pegah Varghaei, Kai Wang, Xingguang Zhang, Stanley Chan, Arun Ross, Humphrey Shi, Zhangyang Wang, Anil Jain, Xiaoming Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6227-6236

Whole-body biometric recognition is an important area of research due to its vast applications in law enforcement, border security, and surveillance. This paper presents the end-to-end design, development and evaluation of FarSight, an innovative software system designed for whole-body (fusion of face, gait and body shape) biometric recognition. FarSight accepts videos from elevated platforms and drones as input and outputs a candidate list of identities from a gallery. The system is designed to address several challenges, including (i) low-quality imagery, (ii) large yaw and pitch angles, (iii) robust feature extraction to accommod ate large intra-person variabilities and large inter-person similarities, and (iv) the large domain gap between training and test sets. FarSight combines the physics of imaging and deep learning models to enhance image restoration and biometric feature encoding. We test FarSight's effectiveness using the newly acquired IARPA Biometric Recognition and Identification at Altitude and Range (BRIAR) dataset. Notably, FarSight demonstrated a substantial performance increase on the BRIAR dataset, with gains of +11.82% Rank-20 identification and +11.3% TAR@1%FAR.

****

Time To Shine: Fine-Tuning Object Detection Models With Synthetic Adverse Weather Images

Thomas Rothmeier, Werner Huber, Alois C. Knoll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4447-4456

The detection of vehicles, pedestrians, and obstacles plays an important role in the decision-making process of autonomous vehicles. While existing methods achieve high detection accuracy under good environmental conditions, they often fail in adverse weather conditions due to limited visibility, blurred contours, and low contrast. These "edge-case" scenarios are not well represented in existing datasets and are not handled properly by object detection algorithms. In our work, we propose a novel approach to synthesising photorealistic and highly diverse scenarios that can be used to fine-tune object detection algorithms in adverse weather conditions such as snow, fog, and rain. The approach uses the Midjourney text-to-image model to create accurate synthetic images of desired weather conditions. Our experiments show that training with our dataset significantly improves detection accuracy in harsh weather conditions. Our results are compared to baseline models and models fine-tuned on augmented clear weather images.

****

Unsupervised and Semi-Supervised Co-Salient Object Detection via Segmentation Frequency Statistics

Souradeep Chakraborty, Shujon Naha, Muhammet Bastan, Amit Kumar K. C., Dimitris Samaras; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 332-342

In this paper, we address the detection of co-occurring salient objects (CoSOD) in an image group using frequency statistics in an unsupervised manner, which further enable us to develop a semi-supervised method. While previous works have mostly focused on fully supervised CoSOD, less attention has been allocated to detecting co-salient objects when limited segmentation annotations are available for training. Our simple yet effective unsupervised method US-CoSOD combines the object co-occurrence frequency statistics of unsupervised single-image semantic segmentations with salient foreground detections using self-supervised feature learning. For the first time, we show that a large unlabeled dataset e.g. ImageNet-1k can be effectively leveraged to significantly improve unsupervised CoSOD performance. Our unsupervised model is a great pre-training initialization for our semi-supervised model SS-CoSOD, especially when very limited labeled data is available for training. To avoid propagating erroneous signals from predictions on unlabeled data, we propose a confidence estimation module to guide our semi-supervised training. Extensive experiments on three CoSOD benchmark datasets show that both of our unsupervised and semi-supervised models outperform the corresponding state-of-the-art models by a significant margin (e.g., on the Cosal2015 dataset, our US-CoSOD model has an 8.8% F-measure gain over a SOTA unsupervised co-segmentation model and our SS-CoSOD model has an 11.81% F-measure gain over a SOTA semi-supervised CoSOD model).

****

3SD: Self-Supervised Saliency Detection With No Labels

Rajeev Yasarla, Renliang Weng, Wongun Choi, Vishal M. Patel, Amir Sadeghian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 313-322

We present a conceptually simple self-supervised method for saliency detection. Our method generates and uses pseudo-ground truth labels for training. The generated pseudo-GT labels don't require any kind of human annotations (e.g., pixel-w ise labels or weak labels like scribbles). Recent works show that features extracted from classification tasks provide important saliency cues like structure and semantic information of salient objects in the image. Our method, called 3SD, exploits this idea by adding a branch for a self-supervised classification task in parallel with salient object detection, to obtain class activation maps (CAM maps). These CAM maps along with the edges of the input image are used to generate the pseudo-GT saliency maps to train our 3SD network. Specifically, we propose a contrastive learning-based training on multiple image patches for the classification task. We show the multi-patch classification with contrastive loss improves the quality of the CAM maps compared to naive classification on the entire image. Experiments on six benchmark datasets demonstrate that without any labels, our 3SD method outperforms all existing weakly supervised and unsupervised methods, and its performance is on par with the fully-supervised methods.

****

Pixel Matching Network for Cross-Domain Few-Shot Segmentation
Hao Chen, Yonghan Dong, Zheming Lu, Yunlong Yu, Jungong Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 978-987

Few-Shot Segmentation (FSS) aims to segment the novel class images with a few annotated samples. In the past, numerous studies have concentrated on cross-category tasks, where the training and testing sets are derived from the same dataset, while these methods face significant difficulties in domain-shift scenarios. To better tackle the cross-domain tasks, we propose a pixel matching network (PMNet) to extract the domain-agnostic pixel-level affinity matching with a frozen backbone and capture both the pixel-to-pixel and pixel-to-patch relations in each support-query pair with the bidirectional 3D convolutions. Different from the existing methods that remove the support background, we design a hysteretic spatial filtering module (HSFM) to filter the background-related query features and retain the foreground-related query features with the assistance of the support background, which is beneficial for eliminating interference objects in the query background. We comprehensively evaluate our PMNet on ten benchmarks under cross-category, cross-dataset, and cross-domain FSS tasks. Experimental results demonstrate that PMNet performs very competitively under different settings with only 0.68M parameters, especially under cross-domain FSS tasks, showing its effectiveness and efficiency.

****

Cross-Domain Few-Shot Incremental Learning for Point-Cloud Recognition
Yuwen Tan, Xiang Xiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2307-2316

Sensing 3D objects is critical when 2D object recognition is not accessible. A robot pre-trained on a large point-cloud dataset will encounter unseen classes of 3D objects after deploying it. Therefore, the robot should be able to learn continuously in real-world scenarios. Few-shot class-incremental learning (FSCIL) requires the model to learn from few-shot new examples continually and not forget past classes. However, there is an implicit but strong assumption in the FSCIL that the distribution of the base and incremental classes is the same. In this paper, we focus on cross-domain FSCIL for point-cloud recognition. We decompose the catastrophic forgetting into base class forgetting and incremental class forgetting and alleviate them separately. We utilize the base model to discriminate base samples and new samples by treating base samples as in-distribution samples, and new objects as out-of-distribution samples. We retain the base model to avoid catastrophic forgetting of base classes and train an extra domain-specific module for all new samples to adapt to new classes. At inference, we first discriminate whether the sample belongs to the base class or the new class. Once classified at the model level, test samples are then passed to the corresponding model for class-level classification. To better mitigate the forgetting of new classes, we adopt the soft label and hard label replay together. Extensive experiments on synthetic-to-real incremental 3D datasets show that our proposed method can balance the performance between the base and new objects and outperforms the previous state-of-the-art methods.

****

Robust Unsupervised Domain Adaptation Through Negative-View Regularization

Joonhyeok Jang, Sunhyeok Lee, Seonghak Kim, Jung-un Kim, Seonghyun Kim, Daeshik Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2462-2471

In the realm of Unsupervised Domain Adaptation (UDA), Vision Transformers (ViTs) have recently demonstrated remarkable adaptability surpassing that of traditional Convolutional Neural Networks (CNNs). Nevertheless, the patch-based structure of ViTs heavily relies on local features within image patches, potentially leading to reduced robustness when confronted with out-of-distribution (OOD) samples. To address this concern, we introduce a novel regularizer tailored specifically for UDA. By leveraging negative views, i.e. target-domain samples applied by negative augmentations, we make the learning process more intricate, thereby preventing models from taking shortcuts in spatial context recognition. We present a novel loss function, rooted in contrastive principles, to effectively distinguish between the negative views and original target samples. By integrating this novel regularizer with existing UDA methodologies, we guide ViTs to prioritize context relationships among local patches, thereby enhancing the robustness of ViTs. Our proposed Negative View-based Contrastive (NVC) regularizer substantially boosts the performance of baseline UDA methods across diverse benchmark datasets. Furthermore, we release new dataset, Retail-71, comprising 71 classes of images commonly encountered in retail stores. Through comprehensive experimentation, we showcase the effectiveness of our approach on traditional benchmarks as well as the novel retail domain. These results substantiate the robust adaptation capabilities of our proposed method. Our method is implemented at our repository.

****

Soft Curriculum for Learning Conditional GANs With Noisy-Labeled and Uncurated Unlabeled Data

Kai Katsumata, Duc Minh Vo, Tatsuya Harada, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5323-5332

Label-noise or curated unlabeled data are used to compensate for the assumption of clean labeled data in training the conditional generative adversarial network; however, satisfying such an extended assumption is occasionally laborious or impractical. As a step towards generative modeling accessible to everyone, we introduce a novel conditional image generation framework that accepts noisy-labeled and uncurated unlabeled data during training: (i) closed-set and open-set label noise in labeled data and (ii) closed-set and open-set unlabeled data. To combat it, we propose soft curriculum learning, which assigns instance-wise weights for adversarial training while assigning new labels for unlabeled data and correcting wrong labels for labeled data. Unlike popular curriculum learning, which uses a threshold to pick the training samples, our soft curriculum controls the effect of each training instance by using the weights predicted by the auxiliary classifier, resulting in the preservation of useful samples while ignoring harmful ones. Our experiments show that our approach outperforms existing semi-supervised and label-noise robust methods in terms of both quantitative and qualitative performance. In particular, the proposed approach matches the performance of (semi-)supervised GANs even with less than half the labeled data.

****

HMP: Hand Motion Priors for Pose and Shape Estimation From Video

Enes Duran, Muhammed Kocabas, Vasileios Choutas, Zicong Fan, Michael J. Black; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6353-6363

Understanding how humans interact with the world necessitates accurate 3D hand pose estimation, a task complicated by the hand's high degree of articulation, frequent occlusions, self-occlusions, and rapid motions. While most existing methods rely on single-image inputs, videos have useful cues to address aforementioned issues. However, existing video-based 3D hand datasets are insufficient for training feedforward models to generalize to in-the-wild scenarios. On the other hand, we have access to large human motion capture datasets which also include ha nd motions, e.g. AMASS. Therefore, we develop a generative motion prior specific for hands, trained on the AMASS dataset which features diverse and high-quality hand motions. This motion prior is then employed for video-based 3D hand motion estimation following a latent optimization approach. Our integration of a robust motion prior significantly enhances performance, especially in occluded scenarios. It produces stable, temporally consistent results that surpass conventional single-frame methods. We demonstrate our method's efficacy via qualitative and quantitative evaluations on the HO3D and DexYCB datasets, with special emphasis on an occlusion-focused subset of HO3D. Code is available at https://hmp.is.tue.mpg.de

****

Amodal Intra-Class Instance Segmentation: Synthetic Datasets and Benchmark

Jiayang Ao, Qiuhong Ke, Krista A. Ehinger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 281-290

Images of realistic scenes often contain intra-class objects that are heavily occluded from each other, making the amodal perception task that requires parsing the occluded parts of the objects challenging. Although important for downstream tasks such as robotic grasping systems, the lack of large-scale amodal datasets with detailed annotations makes it difficult to model intra-class occlusions explicitly. This paper introduces two new amodal datasets for image amodal completion tasks, which contain a total of over 267K images of intra-class occlusion scenarios, annotated with multiple masks, amodal bounding boxes, dual order relations and full appearance for instances and background. We also present a point-supervised scheme with layer priors for amodal instance segmentation specifically designed for intra-class occlusion scenarios. Experiments show that our weakly supervised approach outperforms the SOTA fully supervised methods, while our layer priors design exhibits remarkable performance improvements in the case of intra-class occlusion in both synthetic and real images.

****

RMFER: Semi-Supervised Contrastive Learning for Facial Expression Recognition With Reaction Mashup Video

Yunseong Cho, Chanwoo Kim, Hoseong Cho, Yunhoe Ku, Eunseo Kim, Muhammadjon Boboev, Joonseok Lee, Seungryul Baek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5913-5922

Facial expression recognition (FER) has greatly benefited from deep learning but still faces challenges in dataset collection due to the nuanced nature of facial expressions. In this study, we present a novel unlabeled dataset and semi-supervised contrastive learning framework that utilizes Reaction Mashup (RM) videos, a video that includes multiple individuals reacting to the same film. We created a Reaction Mashup dataset (RMset) from these videos. Our framework integrates three distinct modules: A classification module for supervised facial expression categorization, an attention module for inter-sample attention learning, and a contrastive module for attention-based contrastive learning using RMset. We utilize both the classification and attention modules for the initial training, subsequently incorporating the contrastive module to enhance the learning process. Our experiments demonstrate that our method improves feature learning and outperforms state-of-the-art models on three benchmark FER datasets. Codes are available at https://github.com/yunseongcho/RMFER.

****

AMEND: Adaptive Margin and Expanded Neighborhood for Efficient Generalized Category Discovery

Anwesha Banerjee, Liyana Sahir Kallooriyakath, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2101-2110

Generalized Category Discovery aims to discover and cluster images from previously unseen classes, in addition to classifying images from seen classes correctly. In this work, we propose a simple, yet effective framework for this task, which not only performs on-par or better with the current approaches but is also significantly more efficient in terms of computational requirements. Our first contribution is to use expanded neighborhood information in contrastive learning to generate robust and generalizable features. To generate more discriminative feature representations, especially for fine-grained datasets and confusing classes, we propose a class-wise adaptive margin regularizer that aims at increasing the angular separation among the prototypes of all classes. Extensive experiments on three generic as well as four fine-grained benchmark datasets show the usefulness of the proposed Adaptive Margin and Expanded Neighborhood (AMEND) framework.

****

Brainomaly: Unsupervised Neurologic Disease Detection Utilizing Unannotated T1-Weighted Brain MR Images

Md Mahfuzur Rahman Siddiquee, Jay Shah, Teresa Wu, Catherine Chong, Todd J. Schwedt, Gina Dumkrieger, Simona Nikolova, Baoxin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7573-7582

Harnessing the power of deep neural networks in the medical imaging domain is challenging due to the difficulties in acquiring large annotated datasets, especially for rare diseases, which involve high costs, time, and effort for annotation. Unsupervised disease detection methods, such as anomaly detection, can significantly reduce human effort in these scenarios. While anomaly detection typically focuses on learning from images of healthy subjects only, real-world situations often present unannotated datasets with a mixture of healthy and diseased subjects. Recent studies have demonstrated that utilizing such unannotated images can improve unsupervised disease and anomaly detection. However, these methods do not utilize knowledge specific to registered neuroimages, resulting in a subpar performance in neurologic disease detection. To address this limitation, we propose Brainomaly, a GAN-based image-to-image translation method specifically designed for neurologic disease detection. Brainomaly not only offers tailored image-to-image translation suitable for neuroimages but also leverages unannotated mixed images to achieve superior neurologic disease detection. Additionally, we address the issue of model selection for inference without annotated samples by proposing a pseudo-AUC metric, further enhancing Brainomaly's detection performance. Extensive experiments and ablation studies demonstrate that Brainomaly outperforms existing state-of-the-art unsupervised disease and anomaly detection methods by significant margins in Alzheimer's disease detection using a publicly available dataset and headache detection using an institutional dataset. The code is available from https://github.com/mahfuzmohammad/Brainomaly.

****

Contrastive Learning for Multi-Object Tracking With Transformers

Pierre-François De Plaen, Nicola Marinello, Marc Proesmans, Tinne Tuytelaars, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6867-6877

The DEtection TRansformer (DETR) opened new possibilities for object detection by modeling it as a translation task: converting image features into object-level representations. Previous works typically add expensive modules to DETR to perform Multi-Object Tracking (MOT), resulting in more complicated architectures. We instead show how DETR can be turned into a MOT model by employing an instance-level contrastive loss, a revised sampling strategy and a lightweight assignment method. Our training scheme learns object appearances while preserving detection capabilities and with little overhead. Its performance surpasses the previous state-of-the-art by +2.6 mMOTA on the challenging BDD100K dataset and is comparable to existing transformer-based methods on the MOT17 dataset.

****

BEVMap: Map-Aware BEV Modeling for 3D Perception

Mincheol Chang, Seokha Moon, Reza Mahjourian, Jinkyu Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7419-7428

In autonomous driving applications, there is a strong preference for modeling the world in Bird's-Eye View (BEV), as it leads to improved accuracy and performance. BEV features are widely used in perception tasks since they allow fusing information from multiple views in an efficient manner. However, BEV features generated from camera images are prone to be imprecise due to the difficulty of estimating depth in the perspective view. Improper placement of BEV features limits t he accuracy of downstream tasks. We introduce a method for incorporating map information to improve perspective depth estimation from 2D camera images and thereby producing geometrically- and semantically-robust BEV features. We show that augmenting the camera images with the BEV map and map-to-camera projections can compensate for the depth uncertainty. Experiments on the nuScenes dataset demonstrate that our method outperforms previous approaches using only camera images in segmentation and detection tasks.

****

PreciseDebias: An Automatic Prompt Engineering Approach for Generative AI To Mitigate Image Demographic Biases

Colton Clemmer, Junhua Ding, Yunhe Feng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8596-8605

Recent years have witnessed growing concerns over demographic biases in image-centric applications, including image search engines and generative systems. While the advent of generative AI offers a pathway to mitigate these biases by producing underrepresented images, existing solutions still fail to precisely generate images that reflect specified demographic distributions. In this paper, we propose PreciseDebias, a comprehensive end-to-end framework that can rectify demographic bias in image generation. By leveraging fine-tuned Large Language Models (LLMs) coupled with text-to-image generative models, PreciseDebias transforms generic text prompts to produce images in line with specified demographic distributions. The core component of PreciseDebias is our novel instruction-following LLM, meticulously designed with an emphasis on model bias assessment and balanced model training. Extensive experiments demonstrate the effectiveness of PreciseDebias in rectifying biases pertaining to both ethnicity and gender in images. Furthermore, when compared with two baselines, PreciseDebias illustrates its robustness and capability to capture demographic intricacies. The generalization of PreciseDebias is further illuminated by the diverse images it produces across multiple professions and demographic attributes. To ensure reproducibility, we will make PreciseDebias openly accessible to the broader research community by releasing all models and code.

****

Benchmark Generation Framework With Customizable Distortions for Image Classifier Robustness

Soumyendu Sarkar, Ashwin Ramesh Babu, Sajad Mousavi, Zachariah Carmichael, Vineet Gundecha, Sahand Ghorbanpour, Ricardo Luna Gutierrez, Antonio Guillen, Avisek Naug; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4418-4427

We present a novel framework for generating adversarial benchmarks to evaluate the robustness of image classification models. The RLAB framework allows users to customize the types of distortions to be optimally applied to images, which helps address the specific distortions relevant to their deployment. The benchmark can generate datasets at various distortion levels to assess the robustness of different image classifiers. Our results show that the adversarial samples generated by our framework with any of the image classification models, like ResNet-50, Inception-V3, and VGG-16, are effective and transferable to other models causing them to fail. These failures happen even when these models are adversarially retrained using state-of-the-art techniques, demonstrating the generalizability of our adversarial samples. Our framework also allows the creation of adversarial samples for non-ground truth classes at different levels of intensity, enabling tunable benchmarks for the evaluation of false positives. We achieve competitive performance in terms of net L_2 distortion compared to state-of-the-art benchmark techniques on CIFAR-10 and ImageNet; however, we demonstrate our framework achieves such results with simple distortions like Gaussian noise without introducing unnatural artifacts or color bleeds. This is made possible by a model-based reinforcement learning (RL) agent and a technique that reduces a deep tree search of the image for model sensitivity to perturbations, to a one-level analysis and action. The flexibility of choosing distortions and setting classification probability thresholds for multiple classes makes our framework suitable for algorithmic audits.

****

Shape-Biased CNNs Are Not Always Superior in Out-of-Distribution Robustness

Xinkuan Qiu, Meina Kan, Yongbin Zhou, Yanchao Bi, Shiguang Shan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2326-2335

In recent years, Out-of-Distribution (o.o.d) Robustness has garnered increasing attention in Deep Learning, and shape-biased Convolutional Neural Networks (CNNs) are believed to exhibit higher robustness, attributed to the inherent shape-based decision rule of human cognition. In this work, we delve deeper into the intricate relationship between shape/texture information and o.o.d robustness by leveraging a carefully curated "Category-Balanced ImageNet" dataset. We find that shape information is not always superior in distinguishing distinct categories and shape-biased model is not always superior across various o.o.d scenarios. Motivated by these insightful findings, we design a novel method named Shape-Texture Adaptive Recombination (STAR) to achieve higher o.o.d robustness. A category-balanced dataset is firstly used to pretrain a debiased backbone and three specialized heads, each adept at robustly extracting shape, texture, and debiased features. Subsequently, an instance-adaptive recombination head is trained to adaptively adjust the contributions of these distinctive features for each given instance. Through comprehensive experiments, our proposed method achieves state-of-the-art o.o.d robustness across various scenarios such as image corruptions, adversarial attacks, style shifts, and dataset shifts, demonstrating its effectiveness.

****

Towards Visual Saliency Explanations of Face Verification

Yuhang Lu, Zewei Xu, Touradj Ebrahimi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4726-4735

In the past years, deep convolutional neural networks have been pushing the frontier of face recognition (FR) techniques in both verification and identification scenarios. Despite the high accuracy, they are often criticized for lacking explainability. There has been an increasing demand for understanding the decision-making process of deep face recognition systems. Recent studies have investigated the usage of visual saliency maps as an explanation, but they often lack a discussion and analysis in the context of face recognition. This paper concentrates on explainable face verification tasks and conceives a new explanation framework. Firstly, a definition of the saliency-based explanation method is provided, which focuses on the decisions made by the deep FR model. Secondly, a new model-agnostic explanation method named CorrRISE is proposed to produce saliency maps, which reveal both the similar and dissimilar regions of any given pair of face images. Then, an evaluation methodology is designed to measure the performance of general visual saliency explanation methods in face verification. Finally, substantial visual and quantitative results have shown that the proposed CorrRISE method demonstrates promising results in comparison with other state-of-the-art explainable face verification approaches.

****

Bias and Diversity in Synthetic-Based Face Recognition

Marco Huber, Anh Thi Luu, Fadi Boutros, Arjan Kuijper, Naser Damer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6215-6226

Synthetic data is emerging as a substitute for authentic data to solve ethical and legal challenges in handling authentic face data. The current models can create real-looking face images of people who do not exist. However, it is a known and sensitive problem that face recognition systems are susceptible to bias, i.e. performance differences between different demographic and non-demographics attributes, which can lead to unfair decisions. In this work, we investigate how the diversity of synthetic face recognition datasets compares to authentic datasets, and how the distribution of the training data of the generative models affects the distribution of the synthetic data. To do this, we looked at the distribution of gender, ethnicity, age, and head position. Furthermore, we investigated the concrete bias of three recent synthetic-based face recognition models on the s tudied attributes in comparison to a baseline model trained on authentic data. Our results show that the generator generate a similar distribution as the used training data in terms of the different attributes. With regard to bias, it can be seen that the synthetic-based models share a similar bias behavior with the authentic-based models. However, with the uncovered lower intra-identity attribute consistency seems to be beneficial in reducing bias.

****

Textual Alchemy: CoFormer for Scene Text Understanding

Gayatri Deshmukh, Onkar Susladkar, Dhruv Makwana, Sparsh Mittal, Sai Chandra Teja R.; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2931-2941

The paper presents CoFormer (Convolutional Fourier Transformer), a robust and adaptable transformer architecture designed for a range of scene text tasks. CoFormer integrates convolution and Fourier operations into the transformer architecture. Thus, it leverages convolution properties such as shared weights, local receptive fields, and spatial subsampling, while the Fourier operation emphasizes composite characteristics from the frequency domain. The research further proposes the first pretraining datasets, named Textverse10M-E and Textverse10M-H. Using these datasets, we demonstrate the efficacy of pretraining for scene text understanding. CoFormer achieves state-of-theart results with and without pretraining on two downstream tasks: scene text recognition and scene text style transfer. The paper presents LISTNet (Language Invariant Style Transfer), a novel framework for bi-lingual scene text style transfer. It also introduces three datasets, viz., TST500K for scene text style transfer, CSTR2.5M and Akshara550 for scene text recognition.

****

Data-Centric Debugging: Mitigating Model Failures via Targeted Image Retrieval

Sahil Singla, Atoosa Malemir Chegini, Mazda Moayeri, Soheil Feizi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 63-74

Deep neural networks can be unreliable in the real world when the training set does not adequately cover all the settings where they are deployed. Focusing on image classification, we consider the setting where we have an error distribution E representing a deployment scenario where the model fails. We have access to a small set of samples E_sample from E and it can be expensive to obtain additional samples. In the traditional model development framework, mitigating failures of the model in E can be challenging and is often done in an ad hoc manner. In this paper, we propose a general methodology for model debugging that can systemically improve model performance on E while maintaining its performance on the original test set. Our key assumption is that we have access to a large pool of weakly (noisily) labeled data F. However, naively adding F to the training would hurt model performance due to the large extent of label noise. Our Data-Centric Debugging (DCD) framework carefully creates a debug-train set by selecting images from F that are perceptually similar to the images in E_sample. To do this, we use the l_2 distance in the feature space (penultimate layer activations) of various models including ResNet, Robust ResNet and DINO where we observe DINO ViTs are significantly better at discovering similar images compared to Resnets. Compared to the baselines that maintain model performance on the test set, we achieve significantly (+9.45%) improved results on the debug-heldout sets.

****

DTrOCR: Decoder-Only Transformer for Optical Character Recognition

Masato Fujitake; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8025-8035

Typical text recognition methods rely on an encoder-decoder structure, in which the encoder extracts features from an image, and the decoder produces recognized text from these features. In this study, we propose a simpler and more effective method for text recognition, known as the Decoder-only Transformer for Optical Character Recognition (DTrOCR). This method uses a decoder-only Transformer to take advantage of a generative language model that is pre-trained on a large corpus. We examined whether a generative language model that has been successful in natural language processing can also be effective for text recognition in computer vision. Our experiments demonstrated that DTrOCR outperforms current state-of-the-art methods by a large margin in the recognition of printed, handwritten, and scene text in both English and Chinese.

****

Efficient Transferability Assessment for Selection of Pre-Trained Detectors

Zhao Wang, Aoxue Li, Zhenguo Li, Qi Dou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1679-1689

Large-scale pre-training followed by downstream fine-tuning is an effective solution for transferring deep-learning-based models. Since finetuning all possible pre-trained models is computational costly, we aim to predict the transferability performance of these pre-trained models in a computational efficient manner. Different from previous work that seek out suitable models for downstream classification and segmentation tasks, this paper studies the efficient transferability assessment of pre-trained object detectors. To this end, we build up a detector transferability benchmark which contains a large and diverse zoo of pre-trained detectors with various architectures, source datasets and training schemes. Given this zoo, we adopt 6 target datasets from 5 diverse domains as the downstream target tasks for evaluation. Further, we propose to assess classification and regression sub-tasks simultaneously in a unified framework. Additionally, we design a complementary metric for evaluating tasks with varying objects. Experimental results demonstrate that our method outperforms other state-of-the-art approaches in assessing transferability under different target domains while efficiently reducing wall-clock time 32x and requiring a mere 5.2% memory footprint compared to brute-force fine-tuning of all pre-trained detectors. Our assessment code and benchmark will be publicly available.

****

NVAutoNet: Fast and Accurate 360deg 3D Visual Perception for Self Driving

Trung Pham, Mehran Maghoumi, Wanli Jiang, Bala Siva Sashank Jujjavarapu, Mehdi Sajjadi, Xin Liu, Hsuan-Chu Lin, Bor-Jeng Chen, Giang Truong, Chao Fang, Junghyun Kwon, Minwoo Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7376-7385

Achieving robust and real-time 3D perception is fundamental for autonomous vehicles. While most existing 3D perception methods prioritize detection accuracy, they often overlook critical aspects such as computational efficiency, onboard chip deployment friendliness, resilience to sensor mounting deviations, and adaptability to various vehicle types. To address these challenges, we present NVAutoNet: a specialized Bird's-Eye-View (BEV) perception network tailored explicitly for automated vehicles. NVAutoNet takes synchronized camera images as input and predicts 3D signals like obstacles, freespaces, and parking spaces. The core of NVAutoNet's architecture (image and BEV backbones) relies on efficient convolutional networks, optimized for high performance using TensorRT. Our image-to-BEV transformation employs simple linear layers and BEV look-up tables, ensuring rapid inference speed. Trained on an extensive proprietary dataset, NVAutoNet consistently achieves elevated perception accuracy, operating remarkably at 53 frames per second on the NVIDIA DRIVE Orin SoC. Notably, NVAutoNet demonstrates resilience to sensor mounting deviations arising from diverse car models. Moreover, NVAutoNet excels in adapting to varied vehicle types, facilitated by inexpensive model fine-tuning procedures that expedite compatibility adjustments.

****

VideoFACT: Detecting Video Forgeries Using Attention, Scene Context, and Forensic Traces

Tai D. Nguyen, Shengbang Fang, Matthew C. Stamm; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8563-8573

Fake videos represent an important misinformation threat. While existing forensic networks have demonstrated strong performance on image forgeries, recent results reported on the Adobe VideoSham dataset show that these networks fail to identify fake content in videos. In response, we propose VideoFACT - a new network that is able to detect and localize a wide variety of video forgeries and manipulations. To overcome challenges that existing networks face when analyzing videos

, our network utilizes both forensic embeddings to capture traces left by manipulation, context embeddings to control for variation in forensic traces introduced by video coding, and a deep self-attention mechanism to estimate the quality and relative importance of local forensic embeddings. We create several new video forgery datasets and use these, along with publicly available data, to experimentally evaluate our network's performance. These results show that our proposed network is able to identify a diverse set of video forgeries, including those not encountered during training. Furthermore, we show that our network can be fine-tuned to achieve even stronger performance on challenging AI-based manipulations.

****

TEGLO: High Fidelity Canonical Texture Mapping From Single-View Images

Vishal Vinod, Tanmay Shah, Dmitry Lagun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3585-3595

Recent work in Neural Fields (NFs) learn 3D representations from class-specific single view image collections. However, they are unable to reconstruct the input data preserving high-frequency details. Further, these methods do not disentangle appearance from geometry and hence are not suitable for tasks such as texture transfer and editing. In this work, we propose TEGLO (Textured EG3D-GLO) for learning 3D representations from single view in-the-wild image collections for a given class of objects. We accomplish this by training a conditional Neural Radiance Field (NeRF) without any explicit 3D supervision. We equip our method with editing capabilities by creating a dense correspondence mapping to a 2D canonical space. We demonstrate that such mapping enables texture transfer and texture editing without requiring meshes with shared topology. Our key insight is that by mapping the input image pixels onto the texture space we can achieve near perfect reconstruction (>74 dB PSNR at 1024^2 resolution). Our formulation allows for high quality 3D consistent novel view synthesis with high-frequency details even at megapixel image resolutions.

****

Prototypical Contrastive Network for Imbalanced Aerial Image Segmentation

Keiller Nogueira, Mayara Maezano Faita-Pinheiro, Ana Paula Marques Ramos, Wesley Nunes Gonçalves, José Marcato Junior, Jefersson A. dos Santos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8366-8376

Binary segmentation is the main task underpinning several remote sensing applications, which are particularly interested in identifying and monitoring a specific category/object. Although extremely important, such a task has several challenges, including huge intra-class variance for the background and data imbalance. Furthermore, most works tackling this task partially or completely ignore one or both of these challenges and their developments. In this paper, we propose a novel method to perform imbalanced binary segmentation of remote sensing images based on deep networks, prototypes, and contrastive loss. The proposed approach allows the model to focus on learning the foreground class while alleviating the class imbalance problem by allowing it to concentrate on the most difficult background examples. The results demonstrate that the proposed method outperforms state-of-the-art techniques for imbalanced binary segmentation of remote sensing images while taking much less training time.

****

BoostRad: Enhancing Object Detection by Boosting Radar Reflections

Yuval Haitman, Oded Bialer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1638-1647

Automotive radars have an important role in autonomous driving systems. The main challenge in automotive radar detection is the radar's wide point spread function (PSF) in the angular domain that causes blurriness and clutter in the radar image. Numerous studies suggest employing an 'end-to-end' learning strategy using a Deep Neural Network (DNN) to directly detect objects from radar images. This approach implicitly addresses the PSF's impact on objects of interest. In this paper, we propose an alternative approach, which we term "Boosting Radar Reflections" (BoostRad). In BoostRad, a first DNN is trained to narrow the PSF for all t he reflection points in the scene. The output of the first DNN is a boosted reflection image with higher resolution and reduced clutter, resulting in a sharper and cleaner image. Subsequently, a second DNN is employed to detect objects within the boosted reflection image. We develop a novel method for training the boosting DNN that incorporates domain knowledge of radar's PSF characteristics. BoostRad's performance is evaluated using the RADDet and CARRADA datasets, revealing its superiority over reference methods.

****

Frequency Attention for Knowledge Distillation

Cuong Pham, Van-Anh Nguyen, Trung Le, Dinh Phung, Gustavo Carneiro, Thanh-Toan Do; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2277-2286

Knowledge distillation is an attractive approach for learning compact deep neural networks, which learns a lightweight student model by distilling knowledge from a complex teacher model. Attention-based knowledge distillation is a specific form of intermediate feature-based knowledge distillation that uses attention mechanisms to encourage the student to better mimic the teacher. However, most of the previous attention-based distillation approaches perform attention in the spatial domain, which primarily affects local regions in the input image. This may not be sufficient when we need to capture the broader context or global information necessary for effective knowledge transfer. In frequency domain, since each frequency is determined from all pixels of the image in spatial domain, it can contain global information about the image. Inspired by the benefits of the frequency domain, we propose a novel module that functions as an attention mechanism in the frequency domain. The module consists of a learnable global filter that can adjust the frequencies of student's features under the guidance of the teacher's features, which encourages the student's features to have patterns similar to the teacher's features. We then propose an enhanced knowledge review-based distillation model by leveraging the proposed frequency attention module. The extensive experiments with various teacher and student architectures on image classification and object detection benchmark datasets show that the proposed approach outperforms other knowledge distillation methods.

****

Lost Your Style? Navigating With Semantic-Level Approach for Text-To-Outfit Retrieval

Junkyu Jang, Eugene Hwang, Sung-Hyuk Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8066-8075

Fashion stylists have historically bridged the gap between consumers' desires and perfect outfits, which involve intricate combinations of colors, patterns, and materials. Although recent advancements in fashion recommendation systems have made strides in outfit compatibility prediction and complementary item retrieval, these systems rely heavily on pre-selected customer choices. Therefore, we introduce a groundbreaking approach to fashion recommendations: text-to-outfit retrieval task that generates a complete outfit set based solely on textual descriptions given by users. Our model is devised at three semantic levels--item, style, and outfit--where each level progressively aggregates data to form a coherent outfit recommendation based on textual input. Here, we leverage strategies similar to those in the contrastive language-image pretraining model to address the intricate-style matrix within the outfit sets. Using the Maryland Polyvore and Polyvore Outfit datasets, our approach significantly outperformed state-of-the-art models in text-video retrieval tasks, solidifying its effectiveness in the fashion recommendation domain. This research not only pioneers a new facet of fashion recommendation systems, but also introduces a method that captures the essence of individual style preferences through textual descriptions.

****

MoRF: Mobile Realistic Fullbody Avatars From a Monocular Video

Renat Bashirov, Alexey Larionov, Evgeniya Ustinova, Mikhail Sidorenko, David Svitov, Ilya Zakharkin, Victor Lempitsky; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3545-3555

We present a system to create Mobile Realistic Fullbody (MoRF) avatars. MoRF ava tars are rendered in real-time on mobile devices, learned from monocular videos, and have high realism. We use SMPL-X as a proxy geometry and render it with DNR (neural texture and image-2-image network). We improve on prior work, by overfitting per-frame warping fields in the neural texture space, allowing to better align the training signal between different frames. We also refine SMPL-X mesh fitting procedure to improve the overall avatar quality. In the comparisons to other monocular video-based avatar systems, MoRF avatars achieve higher image sharpness and temporal consistency. Participants of our user study also preferred avatars generated by MoRF.

****

dacl10k: Benchmark for Semantic Bridge Damage Segmentation

Johannes Flotzinger, Philipp J. Rösch, Thomas Braml; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8626-8635

Reliably identifying reinforced concrete defects (RCDs) plays a crucial role in assessing the structural integrity, traffic safety, and long-term durability of concrete bridges, which represent the most common bridge type worldwide. Nevertheless, available datasets for the recognition of RCDs are small in terms of size and class variety, which questions their usability in real-world scenarios and their role as a benchmark. Our contribution to this problem is "dacl10k", an exceptionally diverse RCD dataset for multi-label semantic segmentation comprising 9,920 images deriving from real-world bridge inspections. dacl10k distinguishes 12 damage classes as well as 6 bridge components that play a key role in the building assessment and recommending actions, such as restoration works, traffic load limitations or bridge closures. In addition, we examine baseline models for dacl10k which are subsequently evaluated. The best model achieves a mean intersection-over-union of 0.42 on the test set. dacl10k, along with our baselines, will be openly accessible to researchers and practitioners, representing the currently biggest dataset regarding number of images and class diversity for semantic segmentation in the bridge inspection domain.

****

What's Outside the Intersection? Fine-Grained Error Analysis for Semantic Segmentation Beyond IoU

Maximilian Bernhard, Roberto Amoroso, Yannic Kindermann, Lorenzo Baraldi, Rita Cucchiara, Volker Tresp, Matthias Schubert; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 968-977

Semantic segmentation represents a fundamental task in computer vision with various application areas such as autonomous driving, medical imaging, or remote sensing. For evaluating and comparing semantic segmentation models, the mean intersection over union (mIoU) is currently the gold standard. However, while mIoU serves as a valuable benchmark, it does not offer insights into the types of errors incurred by a model. Moreover, different types of errors may have different impacts on downstream applications. To address this issue, we propose an intuitive method for the systematic categorization of errors, thereby enabling a fine-grained analysis of semantic segmentation models. Since we assign each erroneous pixel to precisely one error type, our method seamlessly extends the popular IoU-based evaluation by shedding more light on the false positive and false negative predictions. Our approach is model- and dataset-agnostic, as it does not rely on additional information besides the predicted and ground-truth segmentation masks. In our experiments, we demonstrate that our method accurately assesses model strengths and weaknesses on a quantitative basis, thus reducing the dependence on time-consuming qualitative model inspection. We analyze a variety of state-of-the-art semantic segmentation models, revealing systematic differences across various architectural paradigms. Exploiting the gained insights, we showcase that combining two models with complementary strengths in a straightforward way is sufficient to consistently improve mIoU, even for models setting the current state of the art on ADE20K. We release a toolkit for our evaluation method at https://github.com/mxbh/beyond-iou.

****

Co-Speech Gesture Detection Through Multi-Phase Sequence Labeling

Esam Ghaleb, Ilya Burenko, Marlou Rasenberg, Wim Pouw, Peter Uhrig, Judith Holler, Ivan Toni, Asl■ Özyürek, Raquel Fernández; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4007-4015

Gestures are integral components of face-to-face communication. They unfold over time, often following predictable movement phases of preparation, stroke, and retraction. Yet, the prevalent approach to automatic gesture detection treats the problem as binary classification, classifying a segment as either containing a gesture or not, thus failing to capture its inherently sequential and contextual nature. To address this, we introduce a novel framework that reframes the task as a multi-phase sequence labeling problem rather than binary classification. Our model processes sequences of skeletal movements over time windows, uses Transformer encoders to learn contextual embeddings, and leverages Conditional Random Fields to perform sequence labeling. We evaluate our proposal on a large dataset of diverse co-speech gestures in task-oriented face-to-face dialogues. The results consistently demonstrate that our method significantly outperforms strong baseline models in detecting gesture strokes. Furthermore, applying Transformer encoders to learn contextual embeddings from movement sequences substantially improves gesture unit detection. These results highlight our framework's capacity to capture the fine-grained dynamics of co-speech gesture phases, paving the way for more nuanced and accurate gesture detection and analysis.

****

Missing Modality Robustness in Semi-Supervised Multi-Modal Semantic Segmentation

Harsh Maheshwari, Yen-Cheng Liu, Zsolt Kira; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1020-1030

Using multiple spatial modalities has been proven helpful in improving semantic segmentation performance. However, there are several real-world challenges that have yet to be addressed: (a) improving label efficiency and (b) enhancing robustness in realistic scenarios where modalities are missing at the test time. To address these challenges, we first propose a simple yet efficient multi-modal fusion mechanism Linear Fusion, that performs better than the state-of-the-art multi-modal models even with limited supervision. Second, we propose M3L: Multi-modal Teacher for Masked Modality Learning, a semi-supervised framework that not only improves the multi-modal performance but also makes the model robust to the realistic missing modality scenario using unlabeled data. We create the first benchmark for semi-supervised multi-modal semantic segmentation and also report the robustness to missing modalities. Our proposal shows an absolute improvement of up to 5% on robust mIoU above the most competitive baselines. Our project page is at https://harshm121.github.io/projects/m3l.html

****

Adversarial Likelihood Estimation With One-Way Flows

Omri Ben-Dov, Pravir Singh Gupta, Victoria Abrevaya, Michael J. Black, Partha Ghosh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3779-3788

Generative Adversarial Networks (GANs) can produce high-quality samples, but do not provide an estimate of the probability density around the samples. However, it has been noted that maximizing the log-likelihood within an energy-based setting can lead to an adversarial framework where the discriminator provides unnormalized density (often called energy). We further develop this perspective, incorporate importance sampling, and show that 1) Wasserstein GAN performs a biased estimate of the partition function, and we propose instead to use an unbiased estimator; and 2) when optimizing for likelihood, one must maximize generator entropy. This is hypothesized to provide a better mode coverage. Different from previous works, we explicitly compute the density of the generated samples. This is the key enabler to designing an unbiased estimator of the partition function and computation of the generator entropy term. The generator density is obtained via a new type of flow network, called one-way flow network, that is less constrained in terms of architecture, as it does not require a tractable inverse function. Our experimental results show that our method converges faster, produces comparable sample quality to GANs with similar architecture, successfully avoids over-fitting to commonly used datasets and produces smooth low-dimensional latent re presentations of the training data.

****

Fast Sun-Aligned Outdoor Scene Relighting Based on TensoRF

Yeonjin Chang, Yearim Kim, Seunghyeon Seo, Jung Yi, Nojun Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3626-3636

In this work, we introduce our method of outdoor scene relighting for Neural Radiance Fields (NeRF) named Sun-aligned Relighting TensoRF (SR-TensoRF). SR-TensoRF offers a lightweight and rapid pipeline aligned with the sun, thereby achieving a simplified workflow that eliminates the need for environment maps. Our sun-alignment strategy is motivated by the insight that shadows, unlike viewpoint-dependent albedo, are determined by light direction. We directly use the sun direction as an input during shadow generation, simplifying the requirements of the inference process significantly. Moreover, SR-TensoRF leverages the training efficiency of TensoRF by incorporating our proposed cubemap concept, resulting in notable acceleration in both training and rendering processes compared to existing methods.

****

Robust Eye Blink Detection Using Dual Embedding Video Vision Transformer

Jeongmin Hong, Joseph Shin, Juhee Choi, Minsam Ko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6374-6384

Eye blink detection serves as a crucial biomarker for evaluating both physical and mental states, garnering considerable attention in biometric and video-based studies. Among various methods, video-based eye blink detection has been particularly favored due to its non-invasive nature, enabling broader applications. However, capturing eye blinks from different camera angles poses significant challenges, primarily because the eye region is relatively small and eye blinks occur rapidly, necessitating a robust detection algorithm. To address these challenges, we introduce Dual Embedding Video Vision Transformer (DE-ViViT), a novel approach for eye blink detection that employs two different embedding strategies: (i) tubelet embedding and (ii) residual embedding. Each embedding can capture large and subtle changes within the eye movement sequence respectively. We rigorously evaluate our proposed method using HUST-LEBW, a publicly available dataset, as well as our newly collected multi-angle eye blink dataset (MAEB). The results indicate that the proposed model consistently outperforms existing methods across both datasets, with notably minor performance variations depending on the camera angles.

****

Domain Generalisation via Risk Distribution Matching

Toan Nguyen, Kien Do, Bao Duong, Thin Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2790-2799

We propose a novel approach for domain generalisation (DG) leveraging risk distributions to characterise domains, thereby achieving domain invariance. In our findings, risk distributions effectively highlight differences between training domains and reveal their inherent complexities. In testing, we may observe similar, or potentially intensifying in magnitude, divergences between risk distributions. Hence, we propose a compelling proposition: Minimising the divergences between risk distributions across training domains leads to robust invariance for DG. The key rationale behind this concept is that a model, trained on domain-invariant or stable features, may consistently produce similar risk distributions across various domains. Building upon this idea, we propose Risk Distribution Matching (RDM). Using the maximum mean discrepancy (MMD) distance, RDM aims to minimise the variance of risk distributions across training domains. However, when the number of domains increases, the direct optimisation of variance leads to linear growth in MMD computations, resulting in inefficiency. Instead, we propose an approximation that requires only one MMD computation, by aligning just two distributions: that of the worst-case domain and the aggregated distribution from all domains. Notably, this method empirically outperforms optimising distributional variance while being computationally more efficient. Unlike conventional DG matching algorithms, RDM stands out for its enhanced efficacy by concentrating on sc alar risk distributions, sidestepping the pitfalls of high-dimensional challenges seen in feature or gradient matching. Our extensive experiments on standard benchmark datasets demonstrate that RDM shows superior generalisation capability over state-of-the-art DG methods.

****

Panelformer: Sewing Pattern Reconstruction From 2D Garment Images

Cheng-Hsiu Chen, Jheng-Wei Su, Min-Chun Hu, Chih-Yuan Yao, Hung-Kuo Chu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 454-463

In this paper, we present a novel approach for reconstructing garment sewing patterns from 2D garment images. Our method addresses the challenge of handling occlusion in 2D images by leveraging the symmetric and correlated nature of garment panels. We introduce a transformer-based deep neural network called Panelformer that learns the parametric space of garment sewing patterns. The network comprises two components: the panel transformer and the stitch predictor. The panel transformer estimates the parametric panel shapes, including the occluded panels, by learning from the visible ones. The stitch predictor determines the stitching information among the predicted panels, enabling the reconstruction of the complete garment. To mitigate the overfitting problem caused by strong panel correlations, we propose two tailor-made data augmentation techniques: panel masking and garment mixing. These techniques generate a wider variety of panel combinations, enhancing the model's robustness and generalization capability. We evaluate the effectiveness of Panelformer using a synthetic dataset with diverse garment types. The experimental results demonstrate that our method outperforms competing baselines and achieves comparable performance to NeuralTailor, which operates on 3D point cloud data. This validates the efficacy of our approach in the context of garment sewing pattern reconstruction. By utilizing 2D images as input, our method expands the potential applications of garment modeling and offers easy accessibility to end users. Our code is available online.

****

Unsupervised Domain Adaptation of MRI Skull-Stripping Trained on Adult Data to Newborns

Abbas Omidi, Aida Mohammadshahi, Neha Gianchandani, Regan King, Lara Leijser, Roberto Souza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7718-7727

Skull-stripping is an important first step when analyzing brain Magnetic Resonance Imaging (MRI) data. Deep learning-based supervised segmentation models, such as the U-net model, have shown promising results in automating this segmentation task. However, when it comes to newborn MRI data, there are no publicly available brain MRI datasets that come with manually annotated segmentation masks to be used as labels during the training of these models. Manual segmentation of brain MR images is time-consuming, labor-intensive, and requires expertise. Furthermore, using a segmentation model trained on adult brain MR images for segmenting newborn brain images is not effective due to a large domain shift between adult and newborn data. As a result, there is a need for more efficient and accurate skull-stripping methods for newborns' brain MRIs. In this paper, we present an unsupervised approach to adapt a U-net skull-stripping model trained on adult MRI to work effectively on newborns. Our results demonstrate the effectiveness of our novel unsupervised approach in enhancing segmentation accuracy. Our proposed method achieved an overall Dice coefficient of 0.916 +- 0.032 (mean +- std), and our ablation studies confirmed the effectiveness of our proposal. Remarkably, despite being unsupervised, our model's performance stands in close proximity to that of the current state-of-the-art supervised models against which we conducted our comparisons. These findings indicate the potential of this method as a valuable, easier, and faster tool for supporting healthcare professionals in the examination of MR images of newborn brains. All the codes are available at: https://github.com/abbasomidi77/DAUnet.

****

Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models

Minxing Zhang, Ning Yu, Rui Wen, Michael Backes, Yang Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4839-4849

Generative models have demonstrated revolutionary success in various visual creation tasks, but in the meantime, they have been exposed to the threat of leaking private information of their training data. Several membership inference attacks (MIAs) have been proposed to exhibit the privacy vulnerability of generative models by classifying a query image as a training dataset member or nonmember. However, these attacks suffer from major limitations, such as requiring shadow models and white-box access, and either ignoring or only focusing on the unique property of diffusion models, which block their generalization to multiple generative models. In contrast, we propose the first generalized membership inference attack against a variety of generative models such as generative adversarial networks, [variational] autoencoders, implicit functions, and the emerging diffusion models. We leverage only generated distributions from target generators and auxiliary non-member datasets, therefore regarding target generators as black boxes and agnostic to their architectures or application scenarios. Experiments validate that all the generative models are vulnerable to our attack. For instance, our work achieves attack AUC > 0.99 against DDPM, DDIM, and FastDPM trained on CIFAR-10 and CelebA. And the attack against VQGAN, LDM (for the text-conditional generation), and LIIF achieves AUC > 0.90. As a result, we appeal to our community to be aware of such privacy leakage risks when designing and publishing generative models.

****

Multitask Vision-Language Prompt Tuning

Sheng Shen, Shijia Yang, Tianjun Zhang, Bohan Zhai, Joseph E. Gonzalez, Kurt Keutzer, Trevor Darrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5656-5667

Prompt Tuning, conditioning on task-specific learned prompt vectors, has emerged as a data-efficient and parameter-efficient method for adapting large pretrained vision-language models to multiple downstream tasks. However, existing approaches usually consider learning prompt vectors for each task independently from scratch, thereby failing to exploit the rich shareable knowledge across different vision-language tasks. In this paper, we propose multitask vision-language prompt tuning (MVLPT), which incorporates cross-task knowledge into prompt tuning for vision-language models. Specifically, (i) we demonstrate the effectiveness of learning a single transferable prompt from multiple source tasks to initialize the prompt for each target task; (ii) we show many target tasks can benefit each other from sharing prompt vectors and thus can be jointly learned via multitask prompt tuning. We benchmark the proposed MVLPT using three representative prompt tuning methods, namely text prompt tuning, visual prompt tuning, and the unified vision-language prompt tuning. Results in 20 vision tasks demonstrate that the proposed approach outperforms all single-task baseline prompt tuning methods, setting the new state-of-the-art on the few-shot ELEVATER benchmarks and cross-task generalization benchmarks. To understand where the cross-task knowledge is most effective, we also conduct a large-scale study on task transferability with 20 vision tasks in 400 combinations for each prompt tuning method. It shows that the most performant MVLPT for each prompt tuning method prefers different task combinations and many tasks can benefit each other, depending on their visual similarity and label similarity.

****

ProcSim: Proxy-Based Confidence for Robust Similarity Learning

Oriol Barbany, Xiaofan Lin, Muhammet Bastan, Arnab Dhua; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1308-1317

Deep Metric Learning (DML) methods aim at learning an embedding space in which distances are closely related to the inherent semantic similarity of the inputs. Previous studies have shown that popular benchmark datasets often contain numerous wrong labels, and DML methods are susceptible to them. Intending to study the effect of realistic noise, we create an ontology of the classes in a dataset an d use it to simulate semantically coherent labeling mistakes. To train robust DML models, we propose ProcSim, a simple framework that assigns a confidence score to each sample using the normalized distance to its class representative. The experimental results show that the proposed method achieves state-of-the-art performance on the DML benchmark datasets injected with uniform and the proposed semantically coherent noise.

****

Hard-Label Based Small Query Black-Box Adversarial Attack

Jeonghwan Park, Paul Miller, Niall McLaughlin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3986-3995

We consider the hard-label based black-box adversarial attack setting which solely observes the target model's predicted class. Most of the attack methods in this setting suffer from impractical number of queries required to achieve a successful attack. One approach to tackle this drawback is utilising the adversarial transferability between white-box surrogate models and black-box target model. However, the majority of the methods adopting this approach are soft-label based to take the full advantage of zeroth-order optimisation. Unlike mainstream methods, we propose a new practical setting of hard-label based attack with an optimisation process guided by a pre-trained surrogate model. Experiments show the proposed method significantly improves the query efficiency of the hard-label based black-box attack across various target model architectures. We find the proposed method achieves approximately 5 times higher attack success rate compared to the benchmarks, especially at the small query budgets as 100 and 250.

****

Learning to Detour: Shortcut Mitigating Augmentation for Weakly Supervised Semantic Segmentation

JuneHyoung Kwon, Eunju Lee, Yunsung Cho, YoungBin Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 819-828

Weakly supervised semantic segmentation (WSSS) employing weak forms of labels has been actively studied to alleviate the annotation cost of acquiring pixel-level labels. However, classifiers trained on biased datasets tend to exploit shortcut features and make predictions based on spurious correlations between certain backgrounds and objects, leading to a poor generalization performance. In this paper, we propose shortcut mitigating augmentation (SMA) for WSSS, which generates synthetic representations of object-background combinations not seen in the training data to reduce the use of shortcut features. Our approach disentangles the object-relevant and background features. We then shuffle and combine the disentangled representations to create synthetic features of diverse object-background combinations. SMA-trained classifier depends less on contexts and focuses more on the target object when making predictions. In addition, we analyzed the behavior of the classifier on shortcut usage after applying our augmentation using an attribution method-based metric. The proposed method achieved the improved performance of semantic segmentation result on PASCAL VOC 2012 and MS COCO 2014 datasets.

****

3D Super-Resolution Model for Vehicle Flow Field Enrichment

Thanh Luan Trinh, Fangge Chen, Takuya Nanri, Kei Akasaka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5826-5835

In vehicle shape design from aerodynamic performance perspective, deep learning methods enable us to estimate the flow field in a short period. However, the estimated flow fields are generally coarse and of low resolution. Therefore, a super-resolution model is required to enrich them. In this study, we propose a novel super-resolution model to enrich the flow fields around the vehicle to a higher resolution. To deal with the complex flow fields of vehicles, we apply the residual-in-residual dense block (RRDB) as the basic network-building unit in the generator without batch normalization. We then apply the relativistic discriminator to provide better feedback regarding the lack of high-frequency components. In addition, we propose a distance-weighted loss to obtain better estimation in wa ke regions and regions near the vehicle surface. Physics-informed loss is used to help the model generate data that satisfies the physical governing equations. We also propose a new training strategy to improve the leaning effectiveness and avoid instability during training. Experimental results demonstrate that the proposed method outperforms the previous study in vehicle flow field enrichment tasks by a significant margin.

****

Multi-View 3D Object Reconstruction and Uncertainty Modelling With Neural Shape Prior

Ziwei Liao, Steven L. Waslander; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3098-3107

3D object reconstruction is important for semantic scene understanding. It is challenging to reconstruct detailed 3D shapes from monocular images directly due to a lack of depth information, occlusion and noise. Most current methods generate deterministic object models without any awareness of the uncertainty of the reconstruction. We tackle this problem by leveraging a neural object representation which learns an object shape distribution from large dataset of 3d object models and maps it into a latent space. We propose a method to model uncertainty as part of the representation and define an uncertainty-aware encoder which generates latent codes with uncertainty directly from individual input images. Further, we propose a method to propagate the uncertainty in the latent code to SDF values and generate a 3d object mesh with local uncertainty for each mesh component. Finally, we propose an incremental fusion method under a Bayesian framework to fuse the latent codes from multi-view observations. We evaluate the system in both synthetic and real datasets to demonstrate the effectiveness of uncertainty-based fusion to improve 3D object reconstruction accuracy.

****

Do VSR Models Generalize Beyond LRS3?

Yasser Abdelaziz Dahou Djilali, Sanath Narayan, Eustache LeBihan, Haithem Boussaid, Ebtesam Almazrouei, Merouane Debbah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6635-6644

The Lip Reading Sentences-3 (LRS3) benchmark has primarily been the focus of intense research in visual speech recognition (VSR) during the last few years. As a result, there is an increased risk of overfitting to its excessively used test set, which is only one hour duration. To alleviate this issue, we build a new VSR test set by closely following the LRS3 dataset creation processes. We then evaluate and analyse the extent to which the current VSR models generalize to the new test data. We evaluate a broad range of publicly available VSR models and find significant drops in performance on our test set, compared to their corresponding LRS3 results. Our results suggest that the increase in word error rates is caused by the models' inability to generalize to slightly "harder" and more realistic lip sequences than those found in the LRS3 test set. Our new test benchmark will be made public in order to enable future research towards more robust VSR models.

****

Context in Human Action Through Motion Complementarity

Eadom Dessalene, Michael Maynord, Cornelia Fermüller, Yiannis Aloimonos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6531-6540

Motivated by Goldman's Theory of Human Action - a framework in which action decomposes into 1) base physical movements, and 2) the context in which they occur - we propose a novel learning formulation for motion and context, where context is derived as the complement to motion. More specifically, we model physical movement through the adoption of Therbligs, a set of elemental physical motions centered around object manipulation. Context is modeled through the use of a contrastive mutual information loss that formulates context information as the action information not contained within movement information. We empirically prove the utility brought by this separation of representation, showing sizable improvements in action recognition and action anticipation accuracies for a variety of models. We present results over two object manipulation datasets: EPIC Kitchens 100, and 50 Salads.

****

D4: Detection of Adversarial Diffusion Deepfakes Using Disjoint Ensembles

Ashish Hooda, Neal Mangaokar, Ryan Feng, Kassem Fawaz, Somesh Jha, Atul Prakash; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3812-3822

Detecting diffusion-generated deepfake images remains an open problem. Current detection methods fail against an adversary who adds imperceptible adversarial perturbations to the deepfake to evade detection. In this work, we propose Disjoint Diffusion Deepfake Detection (D4), a deepfake detector designed to improve black-box adversarial robustness beyond de facto solutions such as adversarial training. D4 uses an ensemble of models over disjoint subsets of the frequency spectrum to significantly improve adversarial robustness. Our key insight is to leverage a redundancy in the frequency domain and apply a saliency partitioning technique to disjointly distribute frequency components across multiple models. We formally prove that these disjoint ensembles lead to a reduction in the dimensionality of the input subspace where adversarial deepfakes lie, thereby making adversarial deepfakes harder to find for black-box attacks. We then empirically validate the D4 method against several black-box attacks and find that D4 significantly outperforms existing state-of-the-art defenses applied to diffusion-generated deepfake detection. We also demonstrate that D4 provides robustness against adversarial deepfakes from unseen data distributions as well as unseen generative techniques.

****

ProS: Facial Omni-Representation Learning via Prototype-Based Self-Distillation

Xing Di, Yiyu Zheng, Xiaoming Liu, Yu Cheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6087-6098

This paper presents a novel approach, called Prototype-based Self-Distillation (ProS), for unsupervised face representation learning. The existing supervised methods heavily rely on a large amount of annotated training facial data, which poses challenges in terms of data collection and privacy concerns. To address these issues, we propose ProS, which leverages a vast collection of unlabeled face images to learn a comprehensive facial omni-representation. In particular, ProS consists of two vision-transformers (teacher and student models) that are trained with different augmented images (cropping, blurring, coloring, etc.). Besides, we build a face-aware retrieval system along with augmentations to obtain the curated images comprising predominantly facial areas. To enhance the discrimination of learned features, we introduce a prototype-based matching loss that aligns the similarity distributions between features (teacher or student) and a set of learnable prototypes. After pre-training, the teacher vision transformer serves as a backbone for downstream tasks, including attribute estimation, expression recognition, and landmark alignment, achieved through simple fine-tuning with additional layers. Extensive experiments demonstrate that our method achieves state-of-the-art performance on various tasks, both in full and few-shot settings. Furthermore, we investigate pre-training with synthetic face images, and ProS exhibits promising performance in this scenario as well.

****

TCP: Triplet Contrastive-Relationship Preserving for Class-Incremental Learning

Shiyao Li, Xuefei Ning, Shanghang Zhang, Lidong Guo, Tianchen Zhao, Huazhong Yang, Yu Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2031-2040

In class-incremental learning (CIL), when deep neural networks learn new classes, their recognition performance in old classes will drop significantly. This phenomenon is widely known as catastrophic forgetting. To alleviate catastrophic forgetting, existing methods store a small portion of old class data with a memory buffer and replay it while learning new classes. These methods suffer from a severe imbalance problem between old and new classes. In this paper, we discover that the imbalance problem in CIL makes it difficult to preserve the feature relation of old classes and hard to learn the feature relation between old and new classes. To mitigate the above two issues, we design a triplet contrastive preser ving (TCP) loss to preserve old knowledge, and propose an asymmetric augmented contrastive learning (A2CL) method to learn new classes. Comprehensive experiments demonstrate the effectiveness of our method, which increases the average accuracies by 1.26% and 0.95% on CIFAR-100 and ImageNet. Especially under smaller memory buffer settings where the imbalance problem is more severe, our method can surpass the baselines by a large margin (up to 3.2%). We also show that TCP can be easily plugged into other methods and further improve their performance.

****

Self-Supervised Learning for Place Representation Generalization Across Appearance Changes

Mohamed Adel Musallam, Vincent Gaudillière, Djamila Aouada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7448-7458

Visual place recognition is a key to unlocking spatial navigation for animals, humans and robots. While state-of-the-art approaches are trained in a supervised manner and, therefore, hardly capture the information needed for generalizing to unusual conditions. We argue that self-supervised learning may help abstracting the place representation so that it can be foreseen, irrespective of the conditions. More precisely, in this paper, we investigate learning features that are robust to appearance modifications while sensitive to geometric transformations in a self-supervised manner. This dual-purpose training is made possible by combining the two self-supervision main paradigms, i.e. contrastive and predictive learning. Our results on standard benchmarks reveal that jointly learning such appearance-robust and geometry-sensitive image descriptors leads to competitive visual place recognition results across adverse seasonal and illumination conditions without requiring any humanannotated labels.

****

Patch-Based Selection and Refinement for Early Object Detection

Tianyi Zhang, Kishore Kasichainula, Yaoxin Zhuo, Baoxin Li, Jae-Sun Seo, Yu Cao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 729-738

Early object detection (OD) is a crucial task for the safety of many dynamic systems. Current OD algorithms have limited success for small objects at a long distance. To improve the accuracy and efficiency of such a task, we propose a novel set of algorithms that divide the image into patches, select patches with objects at various scales, elaborate the details of a small object, and detect it as early as possible. Our approach is built upon a transformer-based network and integrates the diffusion model to improve the detection accuracy. As demonstrated on BDD100K, our algorithms enhance the mAP for small objects from 1.03 to 8.93, and reduce the data volume in computation by more than 77%.

****

Guided Distillation for Semi-Supervised Instance Segmentation

Tariq Berrada, Camille Couprie, Karteek Alahari, Jakob Verbeek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 475-483

Although instance segmentation methods have improved considerably, the dominant paradigm is to rely on fully annotated training images, which are tedious to obtain. To alleviate this reliance, and boost results, semi-supervised approaches leverage unlabeled data as an additional training signal that limits overfitting to the labeled samples. In this context, we present novel design choices to significantly improve teacher-student distillation models. In particular, we (i) improve the distillation approach by introducing a novel "guided burn-in" stage, and (ii) evaluate different instance segmentation architectures, as well as backbone networks and pre-training strategies. Contrary to previous work which uses only supervised data for the burn-in period of the student model, we also use guidance of the teacher model to exploit unlabeled data in the burn-in period. Our improved distillation approach leads to substantial improvements over previous state-of-the-art results. For example, on the Cityscapes dataset we improve mask-AP from 23.7 to 33.9 when using labels for 10% of images, and on the COCO dataset we improve mask-AP from 18.3 to 34.1 when using labels for only 1% of the train ing data.

****

Optimizing Long-Term Robot Tracking With Multi-Platform Sensor Fusion

Giuliano Albanese, Arka Mitra, Jan-Nico Zaech, Yupeng Zhao, Ajad Chhatkuli, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6992-7002

Monitoring a fleet of robots requires stable long-term tracking with re-identification, which is yet an unsolved challenge in many scenarios. One application of this is the analysis of autonomous robotic soccer games at RoboCup. Tracking in these games requires handling of identically looking players, strong occlusions, and non-professional video recordings, but also offers state information estimated by the robots. In order to make effective use of the information coming from the robot sensors, we propose a robust tracking and identification pipeline. It fuses external non-calibrated camera data with the robots' internal states using quadratic optimization for tracklet matching. The approach is validated using game recordings from previous RoboCup World Cup tournaments.

****

HyperMix: Out-of-Distribution Detection and Classification in Few-Shot Settings

Nikhil Mehta, Kevin J. Liang, Jing Huang, Fu-Jen Chu, Li Yin, Tal Hassner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2410-2420

Out-of-distribution (OOD) detection is an important topic for real-world machine learning systems, but settings with limited in-distribution samples have been underexplored. Such few-shot OOD settings are challenging, as models have scarce opportunities to learn the data distribution before being tasked with identifying OOD samples. Indeed, we demonstrate that recent state-of-the-art OOD methods fail to outperform simple baselines in the few-shot setting. We thus propose a hypernetwork framework called HyperMix, using Mixup on the generated classifier parameters, as well as a natural out-of-episode outlier exposure technique that does not require an additional outlier dataset. We conduct experiments on CIFAR-FS and MiniImageNet, significantly outperforming other OOD methods in the few-shot regime.

****

TriPlaneNet: An Encoder for EG3D Inversion

Ananta R. Bhattarai, Matthias Nießner, Artem Sevastopolsky; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3055-3065

Recent progress in NeRF-based GANs has introduced a number of approaches for high-resolution and high-fidelity generative modeling of human heads with a possibility for novel view rendering. At the same time, one must solve an inverse problem to be able to re-render or modify an existing image or video. Despite the success of universal optimization-based methods for 2D GAN inversion, those applied to 3D GANs may fail to extrapolate the result onto the novel view, whereas optimization-based 3D GAN inversion methods are time-consuming and can require at least several minutes per image. Fast encoder-based techniques, such as those developed for StyleGAN, may also be less appealing due to the lack of identity preservation. Our work introduces a fast technique that bridges the gap between the two approaches by directly utilizing the tri-plane representation presented for the EG3D generative model. In particular, we build upon a feed-forward convolutional encoder for the latent code and extend it with a fully-convolutional predictor of tri-plane numerical offsets. The renderings are similar in quality to the ones produced by optimization-based techniques and outperform the ones by encoder-based methods. As we empirically prove, this is a consequence of directly operating in the tri-plane space, not in the GAN parameter space, while making use of an encoder-based trainable approach. Finally, we demonstrate significantly more correct embedding of a face image in 3D than for all the baselines, further strengthened by a probably symmetric prior enabled during training.

****

Elusive Images: Beyond Coarse Analysis for Fine-Grained Recognition

Connor Anderson, Matt Gwilliam, Evelyn Gaskin, Ryan Farrell; Proceedings of the

IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 829-839

While the community has seen many advances in recent years to address the challenging problem of Finegrained Visual Categorization (FGVC), progress seems to be slowing--new state-of-the-art methods often distinguish themselves by improving top-1 accuracy by mere tenths of a percent. However, across all of the now-standard FGVC datasets, there remain sizeable portions of the test data that none of the current state-of-the-art (SOTA) models can successfully predict. This paper provides a framework for identifying and studying the errors that current methods make across diverse fine-grained datasets. Three models of difficulty--Prediction Overlap, Prediction Rank and Pairwise Class Confusion--are employed to highlight the most challenging sets of images and classes. Extensive experiments apply a range of standard and SOTA methods, evaluating them on multiple FGVC domains and datasets. Insights acquired from coupling these difficulty paradigms with the careful analysis of experimental results suggest crucial areas for future FGVC research, focusing critically on the set of elusive images that none of the current models can correctly classify. Code is available at catalys1.github.io/elusive-images-fgvc.

****

Tracking Skiers From the Top to the Bottom

Matteo Dunnhofer, Luca Sordi, Niki Martinel, Christian Micheloni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8511-8521

Skiing is a popular winter sport discipline with a long history of competitive events. In this domain, computer vision has the potential to enhance the understanding of athletes' performance, but its application lags behind other sports due to limited studies and datasets. This paper makes a step forward in filling such gaps. A thorough investigation is performed on the task of skier tracking in a video capturing his/her complete performance. Obtaining continuous and accurate skier localization is preemptive for further higher-level performance analyses. To enable the study, the largest and most annotated dataset for computer vision in skiing, SkiTB, is introduced. Several visual object tracking algorithms, including both established methodologies and a newly introduced skier-optimized baseline algorithm, are tested using the dataset. The results provide valuable insights into the applicability of different tracking methods for vision-based skiing analysis. SkiTB, code, and results are available at https://machinelearning.uniud.it/datasets/skitb.

****

BPKD: Boundary Privileged Knowledge Distillation for Semantic Segmentation

Liyang Liu, Zihan Wang, Minh Hieu Phan, Bowen Zhang, Jinchao Ge, Yifan Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1062-1072

Current knowledge distillation approaches in semantic segmentation tend to adopt a holistic approach that treats all spatial locations equally. However, for dense prediction, students' predictions on edge regions are highly uncertain due to contextual information leakage, requiring higher spatial sensitivity knowledge than the body regions. To address this challenge, this paper proposes a novel approach called boundary-privileged knowledge distillation (BPKD). it distils the knowledge from the teacher model's body and edges separately to the compact student model. Specifically, we employ two distinct loss functions: (i) edge loss, which aims to distinguish between ambiguous classes at the pixel level in edge regions; (ii) body loss, which utilizes shape constraints and selectively attends to the inner-semantic regions. Our experiments demonstrate that the proposed BPKD method provides extensive refinements and aggregation for edge and body regions. Additionally, the method achieves state-of-the-art distillation performance for semantic segmentation on three popular benchmark datasets, highlighting its effectiveness and generalization ability. BPKD shows consistent improvements across a diverse array of lightweight segmentation structures, including both CNNs and transformers, underscoring its architecture-agnostic adaptability.

****

DREAM: Visual Decoding From Reversing Human Visual System

Weihao Xia, Raoul de Charette, Cengiz Oztireli, Jing-Hao Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8226-8235

In this work we present DREAM, an fMRI-to-image method for reconstructing viewed images from brain activities, grounded on fundamental knowledge of the human visual system. We craft reverse pathways that emulate the hierarchical and parallel nature of how humans perceive the visual world. These tailored pathways are specialized to decipher semantics, color, and depth cues from fMRI data, mirroring the forward pathways from visual stimuli to fMRI recordings. To do so, two components mimic the inverse processes within the human visual system: the Reverse Visual Association Cortex (R-VAC) which reverses pathways of this brain region, extracting semantics from fMRI data; the Reverse Parallel PKM (R-PKM) component simultaneously predicting color and depth from fMRI signals. The experiments indicate that our method outperforms the current state-of-the-art models in terms of the consistency of appearance, structure, and semantics. Code will be available at https://github.com/weihaox/DREAM.

****

Seeing Stars: Learned Star Localization for Narrow-Field Astrometry

Violet Felt, Justin Fletcher; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8297-8305

Star localization in astronomical imagery is a computer vision task that underpins satellite tracking. Astronomical star extraction techniques often struggle to detect stars when applied to satellite tracking imagery due to the narrower fields of view and rate track observational modes of satellite tracking telescopes. We present a large dataset of real narrow-field rate-tracked imagery with ground truth stars, created using a combination of existing star detection techniques, an astrometric engine, and a star catalog. We train three state of the art object detection, instance segmentation, and line segment detection models on this dataset and evaluate them with object-wise, pixel-wise, and astrometric metrics. Our proposed approaches require no metadata; when paired with a lost-in-space astrometric engine, they find astrometric fits based solely on uncorrected image pixels. Experimental results on real data indicate the effectiveness of learned star detection: we report astrometric fit rates over double that of classical star detection algorithms, improved dim star recall, and comparable star localization residuals.

****

How Do Deepfakes Move? Motion Magnification for Deepfake Source Detection

Ilke Demir, Umur Aybars Çiftçi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4780-4790

With the proliferation of deep generative models, deepfakes are improving in quality and quantity everyday. However, there are subtle authenticity signals in pristine videos, not replicated by current generative models. We contrast the movement in deepfakes and authentic videos by motion magnification towards building a generalized deepfake source detector. The sub-muscular motion in faces has different interpretations per different generative models, which is reflected in their generative residue. Our approach exploits the difference between real motion and the amplified generative artifacts, by combining deep and traditional motion magnification, to detect whether a video is fake and its source generator if so. Evaluating our approach on two multi-source datasets, we obtain 97.77% and 94.03% for video source detection. Our approach performs at least 4.08% better than the prior deepfake source detector and other complex architectures. We also analyze magnification amount, phase extraction window, backbone network, sample counts, and sample lengths. Finally, we report our results on skin tones and genders to assess the model bias.

****

Separable Self and Mixed Attention Transformers for Efficient Object Tracking

Goutam Yelluru Gopal, Maria A. Amer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6708-6717

The deployment of transformers for visual object tracking has shown state-of-the

-art results on several benchmarks. However, the transformer-based models are under-utilized for Siamese lightweight tracking due to the computational complexity of their attention blocks. This paper proposes an efficient self and mixed attention transformer-based architecture for lightweight tracking. The proposed backbone utilizes the separable mixed attention transformers to fuse the template and search regions during feature extraction to generate superior feature encoding. Our prediction head performs global contextual modeling of the encoded features by leveraging efficient self-attention blocks for robust target state estimation. With these contributions, the proposed lightweight tracker deploys a transformer-based backbone and head module concurrently for the first time. Our ablation study testifies to the effectiveness of the proposed combination of backbone and head modules. Simulations show that our Separable Self and Mixed Attention-based Tracker, SMAT, surpasses the performance of related lightweight trackers on GOT10k, TrackingNet, LaSOT, NfS30, UAV123, and AVisT datasets, while running at 37 fps on CPU, 158 fps on GPU, and having 3.8M parameters. For example, it significantly surpasses the closely related trackers E.T.Track and MixFormerV2-S on GOT10k-test by a margin of 7.9% and 5.8%, respectively, in the AO metric. The tracker code and model is available at https://github.com/goutamyg/SMAT

****

CLIPAG: Towards Generator-Free Text-to-Image Generation

Roy Ganz, Michael Elad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3843-3853

Perceptually Aligned Gradients (PAG) refer to an intriguing property observed in robust image classification models, wherein their input gradients align with human perception and pose semantic meanings. While this phenomenon has gained significant research attention, it was solely studied in the context of unimodal vision-only architectures. In this work, we extend the study of PAG to Vision-Language architectures, which form the foundations for diverse image-text tasks and applications. Through an adversarial robustification finetuning of CLIP, we demonstrate that robust Vision-Language models exhibit PAG in contrast to their vanilla counterparts. This work reveals the merits of CLIP with PAG (CLIPAG) in several vision-language generative tasks. Notably, we show that seamlessly integrating CLIPAG in a "plug-n-play" manner leads to substantial improvements in vision-language generative applications. Furthermore, leveraging its PAG property, CLIPAG enables text-to-image generation without any generative model, which typically requires huge generators.

****

Source-Guided Similarity Preservation for Online Person Re-Identification

Hamza Rami, Jhony H. Giraldo, Nicolas Winckler, Stéphane Lathuilière; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1711-1720

Online Unsupervised Domain Adaptation (OUDA) for person Re-Identification (Re-ID) is the task of continuously adapting a model trained on a well-annotated source-domain dataset to a target domain observed as a data stream. In OUDA, person Re-ID models face two main challenges: catastrophic forgetting and domain shift. In this work, we propose a new Source-guided Similarity Preservation (S2P) framework to alleviate these two problems. Our framework is based on the extraction of a support set composed of source images that maximizes the similarity with the target data. This support set is used to identify feature similarities that must be preserved during the learning process. S2P can incorporate multiple existing UDA methods to mitigate catastrophic forgetting. Our experiments show that S2P outperforms previous state-of-the-art methods on multiple real-to-real and synthetic-to-real challenging OUDA benchmarks.

****

Uncertainty-Weighted Loss Functions for Improved Adversarial Attacks on Semantic Segmentation

Kira Maag, Asja Fischer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3906-3914

State-of-the-art deep neural networks have been shown to be extremely powerful in a variety of perceptual tasks like semantic segmentation. However, these netwo rks are vulnerable to adversarial perturbations of the input which are imperceptible for humans but lead to incorrect predictions. Treating image segmentation as a sum of pixel-wise classifications, adversarial attacks developed for classification models were shown to be applicable to segmentation models as well. In this work, we present simple uncertainty-based weighting schemes for the loss functions of such attacks that (i) put higher weights on pixel classifications which can more easily perturbed and (ii) zero-out the pixel-wise losses corresponding to those pixels that are already confidently misclassified. The weighting schemes can be easily integrated into the loss function of a range of well-known adversarial attackers with minimal additional computational overhead, but lead to significant improved perturbation performance, as we demonstrate in our empirical analysis on several datasets and models.

****

Towards Realistic Generative 3D Face Models

Aashish Rai, Hiresh Gupta, Ayush Pandey, Francisco Vicente Carrasco, Shingo Jason Takagi, Amaury Aubel, Daeil Kim, Aayush Prakash, Fernando De la Torre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3738-3748

In recent years, there has been significant progress in 2D generative face models fueled by applications such as animation, synthetic data generation, and digital avatars. However, due to the absence of 3D information, these 2D models often struggle to accurately disentangle facial attributes like pose, expression, and illumination, limiting their editing capabilities. To address this limitation, this paper proposes a 3D controllable generative face model to produce high-quality albedo and precise 3D shapes by leveraging existing 2D generative models. By combining 2D face generative models with semantic face manipulation, this method enables editing of detailed 3D rendered faces. The proposed framework utilizes an alternating descent optimization approach over shape and albedo. Differentiable rendering is used to train high-quality shapes and albedo without 3D supervision. Moreover, this approach outperforms most state-of-the-art (SOTA) methods in the well-known NoW and REALY benchmarks for 3D face reconstruction. It also outperforms the SOTA reconstruction models in recovering rendered faces' identities across novel poses. Additionally, the paper demonstrates direct control of expressions in 3D faces by exploiting latent space leading to text-based editing of 3D faces.

****

Domain Generalization by Rejecting Extreme Augmentations

Masih Aminbeidokhti, Fidel A. Guerrero Peña, Heitor Rapela Medeiros, Thomas Dubail, Eric Granger, Marco Pedersoli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2215-2225

Data augmentation is one of the most powerful techniques for regularizing deep learning models and improving their recognition performance in a variety of tasks and domains. However, this holds for standard in-domain settings, in which the training and test data follow the same distribution. For the out-domain, in which the test data follows a different and unknown distribution, the best recipe for data augmentation is not clear. In this paper, we show that also for out-domain or domain generalization settings, data augmentation can bring a conspicuous and robust improvement in performance. For doing that, we propose a simple procedure: i) use uniform sampling on standard data augmentation transformations ii) increase transformations strength to adapt to the higher data variance expected when working out of domain iii) devise a new reward function to reject extreme transformations that can harm the training. With this simple formula, our data augmentation scheme achieves comparable or better results to state-of-the-art performance on most domain generalization datasets.

****

Towards Accurate Disease Segmentation in Plant Images: A Comprehensive Dataset Creation and Network Evaluation

Komuravelli Prashanth, Jaladi Sri Harsha, Sivapuram Arun Kumar, Jaladi Srilekha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7086-7094

Automated disease segmentation in plant images plays a crucial role in identifying and mitigating the impact of plant diseases on agricultural productivity. In this study, we address the problem of Northern Leaf Blight (NLB) disease segmentation in maize plants. We present a comprehensive dataset of 1000 plant images annotated with NLB disease regions. We employ the Mask R-CNN and Cascaded Mask R-CNN models with various backbone architectures to perform NLB disease segmentation. The experimental results demonstrate the effectiveness of the models in accurately delineating NLB disease regions. Specifically, the ResNet Strikes Back-50 backbone architecture achieves the highest mean average precision (mAP) score, indicating its ability to capture intricate details of NLB disease spots. Additionally, the cascaded approach enhances segmentation accuracy compared to the single-stage Mask R-CNN models. Our findings provide valuable insights into the performance of different backbone architectures and contribute to the development of automated NLB disease segmentation methods in plant images. The generated dataset and experimental results serve as a resource for further research in plant disease segmentation and management.

****

Deep Subdomain Alignment for Cross-Domain Image Classification

Yewei Zhao, Hu Han, Shiguang Shan, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2820-2829

Unsupervised domain adaptation (UDA), which aims to transfer knowledge learned from a labeled source domain to an unlabeled target domain, is useful for various cross-domain image classification scenarios. A commonly used approach for UDA is to minimize the distribution differences between two domains, and subdomain alignment is found to be an effective method. However, most of the existing subdomain alignment methods are based on adversarial learning and focus on subdomain alignment procedures without considering the discriminability among individual subdomains, resulting in slow convergence and unsatisfactory adaptation results. To address these issues, we propose a novel deep subdomain alignment method for UDA in image classification, which consists of a Union Subdomain Contrastive Learning (USCL) module and a Multi-view Subdomain Alignment (MvSA) strategy. USCL can create discriminative and dispersed subdomains by bringing samples from the same subdomain closer while pushing away samples from different subdomains. MvSA makes use of labeled source domain data and easy target domain data to perform target-to-source and target-to-target alignment. Experimental results on three image classification datasets (Office-31, Office-Home, Visda-17) demonstrate that our proposed method is effective for UDA and achieves promising results in several cross-domain image classification tasks.

****

Classifying Cable Tendency With Semantic Segmentation by Utilizing Real and Simulated RGB Data

Pei-Chun Chien, Powei Liao, Eiji Fukuzawa, Jun Ohya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8430-8438

Cable tendency is the potential shape or characteristic that a cable may possess while being manipulated, of which some are considered erroneous and should be identified as a part of anomaly detection during an automatic manipulation. This research explores the ability of deep-learning models in learning the cable tendencies that, contrary to typical classification tasks of multi-object scenarios, is to differentiate the multiple states displayable by the same object -- in this case, cables. By training multiple models with different combinations of self-collected real-world data and self-generated simulation data, a comparative study is carried out to compare the performance of each approach. In conclusion, the effectiveness of detecting three abnormal states and shapes of cables, and using simulation data is certificated in experiments.

****

Visual Narratives: Large-Scale Hierarchical Classification of Art-Historical Images

Matthias Springstein, Stefanie Schneider, Javad Rahnama, Julian Stalter, Maximilian Kristen, Eric Müller-Budack, Ralph Ewerth; Proceedings of the IEEE/CVF Winte r Conference on Applications of Computer Vision (WACV), 2024, pp. 7220-7230

Iconography refers to the methodical study and interpretation of thematic content in the visual arts, distinguishing it, e.g., from purely formal or aesthetic considerations. In iconographic studies, Iconclass is a widely used taxonomy that encapsulates historical, biblical, and literary themes, among others. However, given the hierarchical nature and inherent complexity of such a taxonomy, it is highly desirable to use automated methods for (Iconclass-based) image classification. Previous studies either focused narrowly on certain subsets of narratives or failed to exploit Iconclass's hierarchical structure. In this paper, we propose a novel approach for Hierarchical Multi-label Classification (HMC) of iconographic concepts in images. We present three strategies, including Large Language Models (LLMs), for the generation of textual image descriptions using keywords extracted from Iconclass. These descriptions are utilized to pre-train a Vision-Language Model (VLM) based on a newly introduced data set of 477,569 images with more than 20,000 Iconclass concepts, far more than considered in previous studies. Furthermore, we present five approaches to multi-label classification, including a novel transformer decoder that leverages hierarchical information from the Iconclass taxonomy. Experimental results show the superiority of this approach over reasonable baselines.

****

Real-Time Weakly Supervised Video Anomaly Detection

Hamza Karim, Keval Doshi, Yasin Yilmaz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6848-6856

Weakly supervised video anomaly detection is an important problem in many real-world applications where during training there are some anomalous videos, in addition to nominal videos, without labelled frames to indicate when the anomaly happens. State-of-the-art methods in this domain typically focus on offline anomaly detection without any concern for real-time detection. Most of these methods rely on ad hoc feature aggregation techniques and the use of metric learning losses, which limit the ability of the models to detect anomalies in real-time. In line with the premise of deep neural networks, there also has been a growing interest in developing end-to-end approaches that can automatically learn effective features directly from the raw data. We propose the first real-time and end-to-end trained algorithm for weakly supervised video anomaly detection. Our training procedure builds upon recent action recognition literature and uses a trainable video model to learn visual features. This is in contrast to existing approaches which largely depend on pre-trained feature extractors. The proposed method significantly improves the anomaly detection speed and AUC performance compared to the existing methods. Specifically, on the UCF-Crime dataset, our method achieves 86.94% AUC with a decision period of 6.4 seconds while the competing methods achieve at most 85.92% AUC with a decision period of 273 seconds.

****

C2AIR: Consolidated Compact Aerial Image Haze Removal

Ashutosh Kulkarni, Shruti S. Phutke, Santosh Kumar Vipparthi, Subrahmanyam Murala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 749-758

Aerial image haze removal deals with improving the visibility and quality of images captured from aerial platforms, such as drones and satellites. Aerial images are commonly used in various applications such as environmental monitoring, and disaster response. These applications usually require cleaner data for accurate functioning. However, atmospheric conditions such as haze or fog can significantly degrade the quality of these images, reducing their contrast, color saturation, and sharpness, making it difficult to extract meaningful information from them. Existing methods rely on computationally heavy and haze density (light, moderate, dense) specific architectures for aerial image dehazing. In light of these limitations, we propose a novel lightweight and consolidated approach for aerial image dehazing. In this approach, we propose Density Aware Query Modulated Block for learning weather degradations in input features and guiding the restoration process. Further, we propose Cross Collaborative Feed-Forward Block for learning to restore varying sizes of the structures in the input images. Finally, we propose Gated Adaptive Feature Fusion block to achieve inter-scale and intra-feature attentive fusion, effective for aerial image restoration. Extensive analysis on benchmark aerial image dehazing datasets and real-world images, along with detailed ablation studies validate the effectiveness of the proposed approach. Further, we have analysed our method for other restoration task such as underwater image enhancement to experiment its wide applicability. The code is available at https: //github.com/AshutoshKulkarni4998/C2AIR.

****

Permutation-Aware Activity Segmentation via Unsupervised Frame-To-Segment Alignment

Quoc-Huy Tran, Ahmed Mehmood, Muhammad Ahmed, Muhammad Naufil, Anas Zafar, Andrey Konin, Zeeshan Zia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6426-6436

This paper presents an unsupervised transformer-based framework for temporal activity segmentation which leverages not only frame-level cues but also segment-level cues. This is in contrast with previous methods which often rely on frame-level information only. Our approach begins with a frame-level prediction module which estimates framewise action classes via a transformer encoder. The frame-level prediction module is trained in an unsupervised manner via temporal optimal transport. To exploit segment-level information, we utilize a segment-level prediction module and a frame-to-segment alignment module. The former includes a transformer decoder for estimating video transcripts, while the latter matches frame-level features with segment-level features, yielding permutation-aware segmentation results. Moreover, inspired by temporal optimal transport, we introduce simple-yet-effective pseudo labels for unsupervised training of the above modules. Our experiments on four public datasets, i.e., 50 Salads, YouTube Instructions, Breakfast, and Desktop Assembly show that our approach achieves comparable or better performance than previous methods in unsupervised activity segmentation.

****

Prototype Learning for Explainable Brain Age Prediction

Linde S. Hesse, Nicola K. Dinsdale, Ana I. L. Namburete; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7903-7913

The lack of explainability of deep learning models limits the adoption of such models in clinical practice. Prototype-based models can provide inherent explainable predictions, but these have predominantly been designed for classification tasks, despite many important tasks in medical imaging being continuous regression problems. Therefore, in this work, we present ExPeRT: an explainable prototype-based model specifically designed for regression tasks. Our proposed model makes a sample prediction from the distances to a set of learned prototypes in latent space, using a weighted mean of prototype labels. The distances in latent space are regularized to be relative to label differences, and each of the prototypes can be visualized as a sample from the training set. The image-level distances are further constructed from patch-level distances, in which the patches of both images are structurally matched using optimal transport. This thus provides an example-based explanation with patch-level detail at inference time. We demonstrate our proposed model for brain age prediction on two imaging datasets: adult MR and fetal ultrasound. Our approach achieved state-of-the-art prediction performance while providing insight into the model's reasoning process.

****

Exploiting CLIP for Zero-Shot HOI Detection Requires Knowledge Distillation at Multiple Levels

Bo Wan, Tinne Tuytelaars; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1805-1815

In this paper, we investigate the task of zero-shot human-object interaction (HOI) detection, a novel paradigm for identifying HOIs without the need for task-specific annotations. To address this challenging task, we employ CLIP, a large-scale pre-trained vision-language model (VLM), for knowledge distillation on multiple levels. To this end, we design a multi-branch neural network that leverages CLIP for learning HOI representations at various levels, including global images

, local union regions encompassing human-object pairs, and individual instances of humans or objects. To train our model, CLIP is utilized to generate HOI scores for both global images and local union regions that serve as supervision signals. The extensive experiments demonstrate the effectiveness of our novel multi-level CLIP knowledge integration strategy. Notably, the model achieves strong performance, which is even comparable with some fully-supervised and weakly-supervised methods on the public HICO-DET benchmark.

****

SDNet: An Extremely Efficient Portrait Matting Model via Self-Distillation

Ziwen Li, Bo Xu, Jiake Xie, Yong Tang, Cheng Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5625-5634

Most existing portrait matting models either require expensive auxiliary information or try to decompose the task into sub-tasks that are usually resource-hungry. These challenges limit its application on low-power computing devices. In addition, mobile networks tend to be less powerful than those cumbersome ones in feature representation mining. In this paper, we propose an extremely efficient portrait matting model via self-distillation (SDNet), that aims to provide a solution to performing accurate and effective portrait matting with limited computing resources. Our SDNet contains only 2M parameters, 2.2% of the parameters of MGM, and 1.5% of that of Matteformer. We introduce the training pipeline of self-distillation that can improve our lightweight baseline model without any parameter addition, network modification, or over-parameterized teacher models which need well-pretraining. Extensive experiments demonstrate the effectiveness of our self-distillation method and the lightweight SDNet network. Our SDNet outperforms the state-of-the-art (SOTA) lightweight approaches on both synthetic and real-world images.

****

Hybrid Neural Diffeomorphic Flow for Shape Representation and Generation via Triplane

Kun Han, Shanlin Sun, Thanh-Tung Le, Xiangyi Yan, Haoyu Ma, Chenyu You, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7707-7717

Deep Implicit Functions (DIFs) have gained popularity in 3D computer vision due to their compactness and continuous representation capabilities. However, addressing dense correspondences and semantic relationships across DIF-encoded shapes remains a critical challenge, limiting their applications in texture transfer and shape analysis. Moreover, recent endeavors in 3D shape generation using DIFs often neglect correspondence and topology preservation. This paper presents HNDF (Hybrid Neural Diffeomorphic Flow), a method that implicitly learns the underlying representation and decomposes intricate dense correspondences into explicitly axis-aligned triplane features. To avoid suboptimal representations trapped in local minima, we propose hybrid supervision that captures both local and global correspondences. Unlike conventional approaches that directly generate new 3D shapes, we further explore the idea of shape generation with deformed template shape via diffeomorphic flows, where the deformation is encoded by the generated triplane features. Leveraging a pre-existing 2D diffusion model, we produce high-quality and diverse 3D diffeomorphic flows through generated triplanes features, ensuring topological consistency with the template shape. Extensive experiments on medical image organ segmentation datasets evaluate the effectiveness of HNDF in 3D shape representation and generation.

****

Volumetric Disentanglement for 3D Scene Manipulation

Sagie Benaim, Frederik Warburg, Peter Ebert Christensen, Serge Belongie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8667-8677

Recently, advances in differential volumetric rendering enabled significant breakthroughs in the photo-realistic and fine-detailed reconstruction of complex 3D scenes, which is key for many virtual reality applications. However, in the context of augmented reality, one may also wish to effect semantic manipulations or augmentations of objects within a scene. To this end, we propose a volumetric fr amework for (i) disentangling or separating, the volumetric representation of a given foreground object from the background, and (ii) semantically manipulating the foreground object, as well as the background. Our method enables the separate control of pixel color and depth as well as 3D similarity transformations of both the foreground and background objects. We subsequently demonstrate our framework's applicability on several downstream manipulation tasks, going beyond the placement and movement of foreground objects. These tasks include object camouflage, non-negative 3D object inpainting, 3D object translation, 3D object inpainting, and 3D text-based object manipulation. Our framework takes as input a set of 2D masks specifying the desired foreground object for training views, together with the associated 2D views and poses, and produces a foreground-background disentanglement that respects the surrounding illumination, reflections, and partial occlusions, which can be applied to both training and novel views.

****

CAILA: Concept-Aware Intra-Layer Adapters for Compositional Zero-Shot Learning

Zhaoheng Zheng, Haidong Zhu, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1721-1731

In this paper, we study the problem of Compositional Zero-Shot Learning (CZSL), which is to recognize novel attribute-object combinations with pre-existing concepts. Recent researchers focus on applying large-scale Vision-Language Pre-trained (VLP) models like CLIP with strong generalization ability. However, these methods treat the pre-trained model as a black box and focus on pre- and post-CLIP operations, which do not inherently mine the semantic concept between the layers inside CLIP. We propose to dive deep into the architecture and insert adapters, a parameter-efficient technique proven to be effective among large language models, into each CLIP encoder layer. We further equip adapters with concept awareness so that concept-specific features of "object", "attribute", and "composition" can be extracted. We assess our method on four popular CZSL datasets, MIT-States, C-GQA, UT-Zappos, and VAW-CZSL, which shows state-of-the-art performance compared to existing methods on all of them.

****

ClusterFix: A Cluster-Based Debiasing Approach Without Protected-Group Supervision

Giacomo Capitani, Federico Bolelli, Angelo Porrello, Simone Calderara, Elisa Ficarra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4870-4879

The failures of Deep Networks can sometimes be ascribed to biases in the data or algorithmic choices. Existing debiasing approaches exploit prior knowledge to avoid unintended solutions; we acknowledge that, in real-world settings, it could be unfeasible to gather enough prior information to characterize the bias, or it could even raise ethical considerations. We hence propose a novel debiasing approach, termed ClusterFix, which does not require any external hint about the nature of biases. Such an approach alters the standard empirical risk minimization and introduces a per-example weight, encoding how critical and far from the majority an example is. Notably, the weights consider how difficult it is for the model to infer the correct pseudo-label, which is obtained in a self-supervised manner by dividing examples into multiple clusters. Extensive experiments show that the misclassification error incurred in identifying the correct cluster allows for identifying examples prone to bias-related issues. As a result, our approach outperforms existing methods on standard benchmarks for bias removal and fairness.

****

Simple Post-Training Robustness Using Test Time Augmentations and Random Forest

Gilad Cohen, Raja Giryes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3996-4006

Although Deep Neural Networks (DNNs) achieve excellent performance on many real-world tasks, they are highly vulnerable to adversarial attacks. A leading defense against such attacks is adversarial training, a technique in which a DNN is trained to be robust to adversarial attacks by introducing adversarial noise to its input. This procedure is effective but must be done during the training phase.

In this work, we propose Augmented Random Forest (ARF), a simple and easy-to-use strategy for robustifying an existing pretrained DNN without modifying its weights. For every image, we generate randomized test time augmentations by applying diverse color, blur, noise, and geometric transforms. Then we use the DNN's logits output to train a simple random forest to predict the real class label. Our method achieves state-of-the-art adversarial robustness on a diversity of white and black box attacks with minimal compromise on the natural images' classification. We test ARF also against numerous adaptive white-box attacks and it shows excellent results when combined with adversarial training.
****

Learning Low-Rank Latent Spaces With Simple Deterministic Autoencoder: Theoretical and Empirical Insights
Alokendu Mazumder, Tirthajit Baruah, Bhartendu Kumar, Rishab Sharma, Vishwajeet Pattanaik, Punit Rathore; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2851-2860
The autoencoder is an unsupervised learning paradigm that aims to create a compact latent representation of data by minimizing the reconstruction loss. However, it tends to overlook the fact that most data (images) are embedded in a lower-dimensional latent space, which is crucial for effective data representation. To address this limitation, we propose a novel approach called Low-Rank Autoencoder (LoRAE). In LoRAE, we incorporated a low-rank regularizer to adaptively learn a low-dimensional latent space while preserving the basic objective of an autoencoder. This helps embed the data in a lower-dimensional latent space while preserving important information. It is a simple autoencoder extension that learns low-rank latent space. Theoretically, we establish a tighter error bound for our model. Empirically, our model's superiority shines through various tasks such as image generation and downstream classification. Both theoretical and practical outcomes highlight the importance of acquiring low-dimensional embeddings.
****

A Hybrid Graph Network for Complex Activity Detection in Video
Salman Khan, Izzeddin Teeti, Andrew Bradley, Mohamed Elhoseiny, Fabio Cuzzolin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6762-6772
Interpretation and understanding of video presents a challenging computer vision task in numerous fields - e.g. autonomous driving and sports analytics. Existing approaches to interpreting the actions taking place within a video clip are based upon Temporal Action Localisation (TAL), which typically identifies short-term actions. The emerging field of Complex Activity Detection (CompAD) extends this analysis to long-term activities, with a deeper understanding obtained by modelling the internal structure of a complex activity taking place within the video. We address the CompAD problem using a hybrid graph neural network which combines attention applied to a graph encoding the local (short-term) dynamic scene with a temporal graph modelling the overall long-duration activity. Our approach is as follows: i) Firstly, we propose a novel feature extraction technique which, for each video snippet, generates spatiotemporal 'tubes' for the active elements ('agents') in the (local) scene by detecting individual objects, tracking them and then extracting 3D features from all the agent tubes as well as the overall scene. ii) Next, we construct a local scene graph where each node (representing either an agent tube or the scene) is connected to all other nodes. Attention is then applied to this graph to obtain an overall representation of the local dynamic scene. iii) Finally, all local scene graph representations are interconnected via a temporal graph, to estimate the complex activity class together with its start and end time. The proposed framework outperforms all previous state-of-the-art methods on all three datasets including ActivityNet-1.3, Thumos-14, and ROAD.
****

Movie Genre Classification by Language Augmentation and Shot Sampling
Zhongping Zhang, Yiwen Gu, Bryan A. Plummer, Xin Miao, Jiayi Liu, Huayan Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7275-7285

Video-based movie genre classification has garnered considerable attention due to its various applications in recommendation systems. Prior work has typically addressed this task by adapting models from traditional video classification tasks, such as action recognition or event detection. However, these models often neglect language elements (e.g., narrations or conversations) present in videos, which can implicitly convey high-level semantics of movie genres, like storylines or background context. Additionally, existing approaches are primarily designed to encode the entire content of the input video, leading to inefficiencies in predicting movie genres. Movie genre prediction may require only a few shots to accurately determine the genres, rendering a comprehensive understanding of the entire video unnecessary. To address these challenges, we propose a Movie genre Classification method based on Language augmentatIon and shot samPling (Movie-CLIP). Movie-CLIP mainly consists of two parts: a language augmentation module to recognize language elements from the input audio, and a shot sampling module to select representative shots from the entire video. We evaluate our method on MovieNet and Condensed Movies datasets, achieving approximate 6-9% improvement in mean Average Precision (mAP) over the baselines. We also generalize Movie-CLIP to the scene boundary detection task, achieving 1.1% improvement in Average Precision (AP) over the state-of-the-art. We release our implementation at github.com/Zhongping-Zhang/Movie-CLIP.

****

Automated Camera Calibration via Homography Estimation With GNNs

Giacomo D'Amicantonio, Egor Bondarev, Peter H.N. de With; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5876-5883

Over the past few decades, a significant rise of camera-based applications for traffic monitoring has occurred. Governments and local administrations are increasingly relying on the data collected from these cameras to enhance road safety and optimize traffic conditions. However, for effective data utilization, it is imperative to ensure accurate and automated calibration of the involved cameras. This paper proposes a novel approach to address this challenge by leveraging the topological structure of intersections. We propose a framework involving the generation of a set of synthetic intersection viewpoint images from a bird's-eye-view image, framed as a graph of virtual cameras to model these images. Using the capabilities of Graph Neural Networks, we effectively learn the relationships within this graph, thereby facilitating the estimation of a homography matrix. This estimation leverages the neighbourhood representation for any real-world camera and is enhanced by exploiting multiple images instead of a single match. In turn, the homography matrix allows the retrieval of extrinsic calibration parameters. As a result, the proposed framework demonstrates superior performance on both synthetic datasets and real-world cameras, setting a new state-of-the-art benchmark.

****

Randomized Adversarial Style Perturbations for Domain Generalization

Taehoon Kim, Bohyung Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2317-2325

We propose a novel domain generalization technique, referred to as Randomized Adversarial Style Perturbation (RASP), which is motivated by the observation that the characteristics of each domain are captured by the feature statistics corresponding to its style. The proposed algorithm perturbs the style of a feature in an adversarial direction towards a randomly selected class, and prevents the model from being misled by the unexpected styles observed in unseen target domains. While RASP is effective for handling domain shifts, its naive integration into the training procedure is prone to degrade the capability of learning knowledge from source domains due to the feature distortions caused by style perturbation. This challenge is alleviated by Normalized Feature Mixup (NFM) during training, which facilitates learning the original features while achieving robustness to perturbed representations. We evaluate the proposed algorithm via extensive experiments on various benchmarks and show that our approach improves domain generalization performance, especially in large-scale benchmarks.

****

C-CLIP: Contrastive Image-Text Encoders To Close the Descriptive-Commentative Gap

William Theisen, Walter J. Scheirer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7241-7250

The interplay between the image and comment on a social media post is one of high importance for understanding its overall message. Recent strides in multimodal embedding models, namely CLIP, have provided an avenue forward in relating image and text. However the current training regime for CLIP models is insufficient for matching content found on social media, regardless of site or language. Current CLIP training data is based on what we call "descriptive" text: text in which an image is merely described. This is something rarely seen on social media, where the vast majority of text content is "commentative" in nature. The captions provide commentary and broader context related to the image, rather than describing what is in it. Current CLIP models perform poorly on retrieval tasks where image-caption pairs display a commentative relationship. Closing this gap would be beneficial for several important application areas related to social media. For instance, it would allow groups focused on Open-Source Intelligence Operations (OSINT) to further aid efforts during disaster events, such as the ongoing Russian invasion of Ukraine, by easily exposing data to non-technical users for discovery and analysis. In order to close this gap we demonstrate that training contrastive image-text encoders on explicitly commentative pairs results in large improvements in retrieval results, with the results extending across a variety of non-English languages.

****

LInKs "Lifting Independent Keypoints" - Partial Pose Lifting for Occlusion Handling With Improved Accuracy in 2D-3D Human Pose Estimation

Peter Hardy, Hansung Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3426-3435

We present LInKs, a novel unsupervised learning method to recover 3D human poses from 2D kinematic skeletons obtained from a single image, even when occlusions are present. Our approach follows a unique two-step process, which involves first lifting the occluded 2D pose to the 3D domain, followed by filling in the occluded parts using the partially reconstructed 3D coordinates. This lift-then-fill approach leads to significantly more accurate results compared to models that complete the pose in 2D space alone. Additionally, we improve the stability and likelihood estimation of normalising flows through a custom sampling function replacing PCA dimensionality reduction used in prior work. Furthermore, we are the first to investigate if different parts of the 2D kinematic skeleton can be lifted independently which we find by itself reduces the error of current lifting approaches. We attribute this to the reduction of long-range keypoint correlations. In our detailed evaluation, we quantify the error under various realistic occlusion scenarios, showcasing the versatility and applicability of our model. Our results consistently demonstrate the superiority of handling all types of occlusions in 3D space when compared to others that complete the pose in 2D space. Our approach also exhibits consistent accuracy in scenarios without occlusion, as evidenced by a 7.9% reduction in reconstruction error compared to prior works on the Human3.6M dataset. Furthermore, our method excels in accurately retrieving complete 3D poses even in the presence of occlusions, making it highly applicable in situations where complete 2D pose information is unavailable.

****

Beyond Classification: Definition and Density-Based Estimation of Calibration in Object Detection

Teodora Popordanoska, Aleksei Tiulpin, Matthew B. Blaschko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 585-594

Despite their impressive predictive performance in various computer vision tasks, deep neural networks (DNNs) tend to make overly confident predictions, which hinders their widespread use in safety-critical applications. While there have been recent attempts to calibrate DNNs, most of these efforts have primarily been focused on classification tasks, thus neglecting DNN-based object detectors. Although several recent works addressed calibration for object detection and proposed differentiable penalties, none of them are consistent estimators of established concepts in calibration. In this work, we tackle the challenge of defining and estimating calibration error specifically for this task. In particular, we adapt the definition of classification calibration error to handle the nuances associated with object detection, and predictions in structured output spaces more generally. Furthermore, we propose a consistent and differentiable estimator of the detection calibration error, utilizing kernel density estimation. Our experiments demonstrate the effectiveness of our estimator against competing train-time and post-hoc calibration methods, while maintaining similar detection performance.

****

PrivObfNet: A Weakly Supervised Semantic Segmentation Model for Data Protection

ChiatPin Tay, Vigneshwaran Subbaraju, Thivya Kandappu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2421-2431

The use of social media has made it easy to communicate and share information over the internet. However, it also brings issues such as data privacy leakage, which can be exploited by recipients with malicious intentions to harm the sender. In this paper, we propose a deep neural network that analyzes the user's image for privacy sensitive content and automatically locates sensitive regions for obfuscation. Our approach relies solely on image level annotations and learns to (a) predict an overall privacy score, (b) detect sensitive attributes and (c) demarcate the sensitive regions for obfuscation, in a given input image. We validated the performance of our proposed method on three large datasets, VISPR, PASCAL VOC 2012 and MS COCO 2014, in terms of privacy score, attribute prediction and obfuscation performance. On the VISPR dataset, we achieved a Pearson correlation of 0.88 and a Spearman correlation of 0.86, outperforming previous methods. On PASCAL VOC 2012 and MS COCO 2014, our model achieved a mean IOU of 71.5% and 43.9% respectively, and is among the state-of-the-art techniques using weakly supervised semantic segmentation learning.

****

Toward Planet-Wide Traffic Camera Calibration

Khiem Vuong, Robert Tamburo, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8553-8562

Despite the widespread deployment of outdoor cameras, their potential for automated analysis remains largely untapped due, in part, to calibration challenges. The absence of precise camera calibration data, including intrinsic and extrinsic parameters, hinders accurate real-world distance measurements from captured videos. To address this, we present a scalable framework that utilizes street-level imagery to reconstruct a metric 3D model, facilitating precise calibration of in-the-wild traffic cameras. Notably, our framework achieves 3D scene reconstruction and accurate localization of over 100 global traffic cameras and is scalable to any camera with sufficient street-level imagery. For evaluation, we introduce a dataset of 20 fully calibrated traffic cameras, demonstrating our method's significant enhancements over existing automatic calibration techniques. Furthermore, we highlight our approach's utility in traffic analysis by extracting insights via 3D vehicle reconstruction and speed measurement, thereby opening up the potential of using outdoor cameras for automated analysis.

****

3D Human Pose Estimation With Two-Step Mixed-Training Strategy

Yingfeng Wang, Zhengwei Wang, Muyu Li, Hong Yan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3332-3341

In monocular 3D human pose estimation, target motions are generally stable and continuous, which indicates that joint velocity can provide valuable information for better estimation. Therefore, it is critical to learn the joint motion trajectory and spatio-temporal information from velocity. Previous works have shown that Transformers are effective in capturing the relationship between tokens. How ever, in practice, only 2D position is available and 3D velocity has not been explicitly used as a model input. To address this challenge, we propose TMT (Two-step Mixed-Training strategy), a transformer-based approach that effectively incorporates 3D velocity into the input vector during training, allowing for better learning of relevant features in the shallow layers. Extensive experiments demonstrate that TMT significantly improves the performance of state-of-the-art models, such as MixSTE, MHFormer, and PoseFomer, on two datasets: Human3.6M and MPI-INF-3DHP. TMT out performs the state-of-the-art approach by up to 13.8% on the Human3.6M dataset.

****

Learning-Based Spotlight Position Optimization for Non-Line-of-Sight Human Localization and Posture Classification

Sreenithy Chandran, Tatsuya Yatagawa, Hiroyuki Kubo, Suren Jayasuriya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4218-4227

Non-line-of-sight imaging (NLOS) is the process of estimating information about a scene that is hidden from the direct line of sight of the camera. NLOS imaging typically requires time-resolved detectors and a laser source for illumination, which are both expensive and computationally intensive to handle. In this paper, we propose an NLOS-based localization and posture classification technique that works on a system of an off-the-shelf projector and camera. We leverage a message-passing neural network to learn a scene geometry and predict the best position to be spotlighted by the projector that can maximize the NLOS signal. The training of the neural network is performed in an end-to-end manner. Therefore, the ground truth spotlighted position is unnecessary during the training, and the network parameters are optimized to maximize the NLOS performance. Unlike prior deep-learning-based NLOS techniques that assume planar relay walls, our system allows us to handle line-of-sight scenes where scene geometries are more arbitrary. Our method demonstrates state-of-the-art performance in object localization and position classification using both synthetic and real scenes.

****

Generalization by Adaptation: Diffusion-Based Domain Extension for Domain-Generalized Semantic Segmentation

Joshua Niemeijer, Manuel Schwonberg, Jan-Aike Termöhlen, Nico M. Schmidt, Tim Fingscheidt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2830-2840

When models, e.g., for semantic segmentation, are applied to images that are vastly different from training data, the performance will drop significantly. Domain adaptation methods try to overcome this issue, but need samples from the target domain. However, this might not always be feasible for various reasons and therefore domain generalization methods are useful as they do not require any target data. We present a new diffusion-based domain extension (DIDEX) method and employ a diffusion model to generate a pseudo-target domain with diverse text prompts. In contrast to existing methods, this allows to control the style and content of the generated images and to introduce a high diversity. In a second step, we train a generalizing model by adapting towards this pseudo-target domain. We outperform previous approaches by a large margin across various datasets and architectures without using any real data. For the generalization from GTA5, we improve state-of-the-art mIoU performance by 3.8% absolute on average and for SYNTHIA by 11.8% absolute, marking a big step for the generalization performance on these benchmarks. Code is available at https://github.com/JNiemeijer/DIDEX

****

Temporally-Consistent Video Semantic Segmentation With Bidirectional Occlusion-Guided Feature Propagation

Razieh Kaviani Baghbaderani, Yuanxin Li, Shuangquan Wang, Hairong Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 685-695

Despite recent progress in static image segmentation, video segmentation is still challenging due to the need for an accurate, fast, and temporally consistent model. Conducting per-frame static image segmentation is not acceptable since it

is computationally prohibitive and prone to temporal inconsistency. In this paper, we present bidirectional occlusion-guided feature propagation (BOFP) method with the goal of improving temporal consistency of segmentation results without sacrificing segmentation accuracy, while at the same time keeping the operations at a low computation cost. It leverages temporal coherence in the video by feature propagation from keyframes to other frames along the motion paths in both forward and backward directions. We propose an occlusion-based attention network to estimate the distorted areas based on bidirectional optical flows, and utilize them as cues for correcting and fusing the propagated features. Extensive experiments on benchmark datasets demonstrate that the proposed BOFP method achieves superior performance in terms of temporal consistency while maintaining comparable level of segmentation accuracy at a low computation cost, striking a great balance among the three metrics essential to evaluate video segmentation solutions.
****
MICS: Midpoint Interpolation To Learn Compact and Separated Representations for Few-Shot Class-Incremental Learning
Solang Kim, Yuho Jeong, Joon Sung Park, Sung Whan Yoon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2236-2245
Few-shot class-incremental learning (FSCIL) aims to learn a classification model for continually accepting novel classes with a few samples. The key of FSCIL is the joint success of the following two training stages: Base training stage to classify base classes and Incremental training stage with sequential learning of novel classes. However, recent efforts show a tendency to focus on one of the stages, or separately design strategies for each stage, so that less effort has been paid to devise a consistent strategy across the consecutive stages. In this paper, we first emphasize the particular aspects of the successful FSCIL algorithm that are worthwhile to consistently pursue during both stages, i.e., intra-class compactness and inter-class separability of the representation, which allows a model to reserve feature space in between current classes for preparing the acceptance of novel classes in the future. To achieve these aspects, we propose a mixup-based FSCIL method called MICS, which theoretically guarantees to enlarge the thickness of the margin space between different classes, leading to outstanding performance on the existing benchmarks. Code is available at https://github.com/solangii/MICS.
****
ParticleNeRF: A Particle-Based Encoding for Online Neural Radiance Fields
Jad Abou-Chakra, Feras Dayoub, Niko Sünderhauf; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5975-5984
While existing Neural Radiance Fields (NeRFs) for dynamic scenes are offline methods with an emphasis on visual fidelity, our paper addresses the online use case that prioritises real-time adaptability. We present ParticleNeRF, a new approach that dynamically adapts to changes in the scene geometry by learning an up-to-date representation online, every 200ms. ParticleNeRF achieves this using a novel particle-based parametric encoding. We couple features to particles in space and backpropagate the photometric reconstruction loss into the particles' position gradients, which are then interpreted as velocity vectors. Governed by a lightweight physics system to handle collisions, this lets the features move freely with the changing scene geometry. We demonstrate ParticleNeRF on various dynamic scenes containing translating, rotating, articulated, and deformable objects. ParticleNeRF is the first online dynamic NeRF and achieves fast adaptability with better visual fidelity than brute-force online InstantNGP and other baseline approaches on dynamic scenes with online constraints.
****
Residual Graph Convolutional Network for Bird's-Eye-View Semantic Segmentation
Qiuxiao Chen, Xiaojun Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3324-3331
Retrieving spatial information and understanding the semantic information of the surroundings are important for Bird's-Eye-View (BEV) semantic segmentation. In the application of autonomous driving, autonomous vehicles need to be aware of t

heir surroundings to drive safely. However, current BEV semantic segmentation techniques, deep Convolutional Neural Networks (CNNs) and transformers, have difficulties in efficiently obtaining the global semantic relationships of the surroundings. In this paper, we propose to incorporate a novel Residual Graph Convolutional (RGC) module in deep CNNs to acquire both the global information and the region-level semantic relationship in the multi-view image domain. Specifically, the RGC module employs a non-overlapping graph space projection to efficiently project the complete BEV information into graph space. It then builds interconnected spatial and channel graphs to extract spatial information between each node and channel information within each node (i.e., extract contextual relationships of the global features). Furthermore, it uses a downsample residual process to enhance the coordinate feature reuse to maintain the global information. The segmentation data augmentation and alignment module helps to simultaneously augment and align BEV features and ground truth to geometrically preserve their alignment to achieve better segmentation results. Our experimental results on the nuScenes benchmark dataset demonstrate that the RGC network outperforms four state-of-the-art networks and its four variants in terms of IoU and mIoU. The proposed RGC network achieves a higher mIoU of 3.1% than the best state-of-the-art network, BEVFusion. Code and models will be released.
****

Group-Wise Contrastive Bottleneck for Weakly-Supervised Visual Representation Learning

Boon Peng Yap, Beng Koon Ng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2246-2255

Coarse or weak labels can serve as a cost-effective solution to the problem of visual representation learning. When fine-grained labels are unavailable, weak labels can provide some form of supervisory signals to guide the representation learning process. Some examples of weak labels include image captions, visual attributes and coarse-grained object categories. In this work, we consider the semantic grouping relationship that exists within certain types of weak labels and propose a group-wise contrastive bottleneck module to leverage this relationship. The semantic group may contain labels that are related to a general concept, such as the colour or shape of objects. Using the group-wise bottleneck module, we disentangle the global image features into multiple group features and apply contrastive learning in a group-wise manner to maximize the similarity of positive pairs within each semantic group. The positive pairs are defined based on the similarity of the labels captured by each group. To learn a more robust representation, we introduce a reconstruction objective where an image feature is reconstructed back from the disentangled features, and this reconstruction is encouraged to be consistent with the feature obtained from a different augmented view of the same image. We empirically verify the efficacy of the proposed method on several datasets in the context of visual attribute learning, fair representation learning and hierarchical label learning. The experimental results indicate that our proposed method outperforms prior weakly-supervised methods and is flexible in adapting to different representation learning settings.
****

Leveraging Synthetic Data To Learn Video Stabilization Under Adverse Conditions

Abdulrahman Kerim, Washington L. S. Ramos, Leandro Soriano Marcolino, Erickson R. Nascimento, Richard Jiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6931-6940

Stabilization plays a central role in improving the quality of videos. However, current methods perform poorly under adverse conditions. In this paper, we propose a synthetic-aware adverse weather video stabilization algorithm that dispenses real data for training, relying solely on synthetic data. Our approach leverages specially generated synthetic data to avoid the feature extraction issues faced by current methods. To achieve this, we present a novel data generator to produce the required training data with an automatic ground-truth extraction procedure. We also propose a new dataset, VSAC105Real, and compare our method to five recent video stabilization algorithms using two benchmarks. Our method generalizes well on real-world videos across all weather conditions and does not require large-scale synthetic training data. Implementations for our proposed video stabilization algorithm, generator, and datasets are available at https://github.com/A-Kerim/SyntheticData4VideoStabilization_WACV_2024.

****

Generation of Upright Panoramic Image From Non-Upright Panoramic Image

Jingguo Liu, Heyu Chen, Shigang Li, Jianfeng Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5261-5270

The inclination of a spherical camera results in nonupright panoramic images. To carry out upright adjustment, traditional methods estimate camera inclination angles firstly, and then resample the image in terms of the estimated rotation to generate upright image. Since sampling an image is a time-consuming processing, a lookup table is usually used to achieve a high processing speed; however, the content of a lookup table depends on the rotational angles and needs extra memory to store also. In this paper we propose a new approach for panorama upright adjustment, which directly generates an upright panoramic image from an input nonupright one without rotation estimation and lookup tables as an intermediate processing. The proposed approach formulates panorama upright adjustment as a pixelwise image-to-image mapping problem, and the mapping is directly generated from an input nonupright panoramic image via an end-to-end neural network. As shown in the experiment of this paper, the proposed method results in a lightweight network, as less as 163MB, with high processing speed, as great as 9ms, for a 256x512 pixel panoramic image.

****

RADIO: Reference-Agnostic Dubbing Video Synthesis

Dongyeun Lee, Chaewon Kim, Sangjoon Yu, Jaejun Yoo, Gyeong-Moon Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4168-4178

One of the most challenging problems in audio-driven talking head generation is achieving high-fidelity detail while ensuring precise synchronization. Given only a single reference image, extracting meaningful identity attributes becomes even more challenging, often causing the network to mirror the facial and lip structures too closely. To address these issues, we introduce RADIO, a framework engineered to yield high-quality dubbed videos regardless of the pose or expression in reference images. The key is to modulate the decoder layers using latent space composed of audio and reference features. Additionally, we incorporate ViT blocks into the decoder to emphasize high-fidelity details, especially in the lip region. Our experimental results demonstrate that RADIO displays high synchronization without the loss of fidelity. Especially in harsh scenarios where the reference frame deviates significantly from the ground truth, our method outperforms state-of-the-art methods, highlighting its robustness.

****

A Coarse-To-Fine Pseudo-Labeling (C2FPL) Framework for Unsupervised Video Anomaly Detection

Anas Al-lahham, Nurbek Tastan, Muhammad Zaigham Zaheer, Karthik Nandakumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6793-6802

Detection of anomalous events in videos is an important problem in applications such as surveillance. Video anomaly detection (VAD) is well-studied in the one-class classification (OCC) and weakly supervised (WS) settings. However, fully unsupervised (US) video anomaly detection methods, which learn a complete system without any annotation or human supervision, have not been explored in depth. This is because the lack of any ground truth annotations significantly increases the magnitude of the VAD challenge. To address this challenge, we propose a simple-but-effective two-stage pseudo-label generation framework that produces segment-level (normal/anomaly) pseudo-labels, which can be further used to train a segment-level anomaly detector in a supervised manner. The proposed coarse-to-fine pseudo-label (C2FPL) generator employs carefully-designed hierarchical divisive clustering and statistical hypothesis testing to identify anomalous video segments from a set of completely unlabeled videos. The trained anomaly detector can be directly applied on segments of an unseen test video to obtain segment-level, a

nd subsequently, frame-level anomaly predictions. Extensive studies on two large-scale public-domain datasets, UCF-Crime and XD-Violence, demonstrate that the proposed unsupervised approach achieves superior performance compared to all existing OCC and US methods, while yielding comparable performance to the state-of-the-art WS methods.
****

Occlusion Sensitivity Analysis With Augmentation Subspace Perturbation in Deep Feature Space
Pedro H. V. Valois, Koichiro Niinuma, Kazuhiro Fukui; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4829-4838

Deep Learning of neural networks has gained prominence in multiple life-critical applications like medical diagnoses and autonomous vehicle accident investigations. However, concerns about model transparency and biases persist. Explainable methods are viewed as the solution to address these challenges. In this study, we introduce the Occlusion Sensitivity Analysis with Deep Feature Augmentation Subspace (OSA-DAS), a novel perturbation-based interpretability approach for computer vision. While traditional perturbation methods make only use of occlusions to explain the model predictions, OSA-DAS extends standard occlusion sensitivity analysis by enabling the integration with diverse image augmentations. Distinctly, our method utilizes the output vector of a DNN to build low-dimensional subspaces within the deep feature vector space, offering a more precise explanation of the model prediction. The structural similarity between these subspaces encompasses the influence of diverse augmentations and occlusions. We test extensively on the ImageNet-1k, and our class- and model-agnostic approach outperforms commonly used interpreters, setting it apart in the realm of explainable AI.
****

PhISH-Net: Physics Inspired System for High Resolution Underwater Image Enhancement
Aditya Chandrasekar, Manogna Sreenivas, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1506-1516

Underwater imaging presents numerous challenges due to refraction, light absorption, and scattering, resulting in color degradation, low contrast, and blurriness. Enhancing underwater images is crucial for high-level computer vision tasks, but existing methods either neglect the physics-based image formation process or require expensive computations. In this paper, we propose an effective framework that combines a physics-based Underwater Image Formation Model (UIFM) with a deep image enhancement approach based on the retinex model. Firstly, we remove backscatter by estimating attenuation coefficients using depth information. Then, we employ a retinex model-based deep image enhancement module to enhance the images. To ensure adherence to the UIFM, we introduce a novel Wideband Attenuation prior. The proposed PhISH-Net framework achieves real-time processing of high-resolution underwater images using a lightweight neural network and a bilateral-grid-based upsampler. Extensive experiments on two underwater image datasets demonstrate the superior performance of our method compared to state-of-the-art techniques. Additionally, qualitative evaluation on a cross-dataset scenario confirms its generalization capability. Our contributions lie in combining the physics-based UIFM with deep image enhancement methods, introducing the wideband attenuation prior, and achieving superior performance and efficiency.
****

MixtureGrowth: Growing Neural Networks by Recombining Learned Parameters
Chau Pham, Piotr Teterwak, Soren Nelson, Bryan A. Plummer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2800-2809

Most deep neural networks are trained under fixed network architectures and require retraining when the architecture changes. If expanding the network's size is needed, it is necessary to retrain from scratch, which is expensive. To avoid this, one can grow from a small network by adding random weights over time to gradually achieve the target network size. However, this naive approach falls short

in practice as it brings too much noise to the growing process. Prior work tackled this issue by leveraging the already learned weights and training data for generating new weights through conducting a computationally expensive analysis step. In this paper, we introduce MixtureGrowth, a new approach to growing networks that circumvents the initialization overhead in prior work. Before growing, each layer in our model is generated with a linear combination of parameter templates. Newly grown layer weights are generated by using a new linear combination of existing templates for a layer. On one hand, these templates are already trained for the task, providing a strong initialization. On the other, the new coefficients provide flexibility for the added layer weights to learn something new. We show that our approach boosts top-1 accuracy over the state-of-the-art by 2-2.5% on CIFAR-100 and ImageNet datasets, while achieving comparable performance with fewer FLOPs to a larger network trained from scratch. Code is available at https://github.com/chaudatascience/mixturegrowth
****

Zero-Shot Building Attribute Extraction From Large-Scale Vision and Language Models

Fei Pan, Sangryul Jeon, Brian Wang, Frank Mckenna, Stella X. Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8647-8656

Modern building recognition methods, exemplified by the BRAILS framework, utilize supervised learning to extract information from satellite and street-view images for image classification and semantic segmentation tasks. However, each task module requires human-annotated data, hindering the scalability and robustness to regional variations and annotation imbalances. In response, we propose a new zero-shot workflow for building attribute extraction that utilizes large-scale vision and language models to mitigate reliance on external annotations. The proposed workflow contains two key components: image-level captioning and segment-level captioning for the building images based on the vocabularies pertinent to structural and civil engineering. These two components generate descriptive captions by computing feature representations of the image and the vocabularies, and facilitating a semantic match between the visual and textual representations. Consequently, our framework offers a promising avenue to enhance AI-driven captioning for building attribute extraction in the structural and civil engineering domains, ultimately reducing reliance on human annotations while bolstering performance and adaptability.
****

SimA: Simple Softmax-Free Attention for Vision Transformers

Soroush Abbasi Koohpayegani, Hamed Pirsiavash; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2607-2617

Recently, vision transformers have become very popular. However, deploying them in many applications is computationally expensive partly due to the Softmax layer in the attention block. We introduce a simple yet effective, Softmax-free attention block, SimA, which normalizes query and key matrices with simple l1-norm instead of using Softmax layer. Then, the attention block in SimA is a simple multiplication of three matrices, so SimA can dynamically change the ordering of the computation at the test time to achieve linear computation on the number of tokens or the number of channels. We empirically show that SimA applied to three SOTA variations of transformers, DeiT, XCiT, and CvT, results in on-par accuracy compared to the SOTA models, without any need for Softmax layer. Interestingly, changing SimA from multi-head to single-head has only a small effect on the accuracy, which further simplifies the attention block. Moreover, we show that SimA is much faster on small edge devices, e.g., Raspberry Pi, which we believe is due to higher complexity of Softmax layer on those devices. The code is available here: https://github.com/UCDvision/sima
****

POP-VQA - Privacy Preserving, On-Device, Personalized Visual Question Answering

Pragya Paramita Sahu, Abhishek Raut, Jagdish Singh Samant, Mahesh Gorijala, Vignesh Lakshminarayanan, Pinaki Bhaskar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8470-8479

The next generation of device smartness needs to go beyond being able to understand basic user commands. As our systems become more efficient, they need to be taught to understand user interactions and intents from all possible input modalities. This is where the recent advent of large scale multi-modal models can form the foundation for next-gen technologies. However, the true power of such interactive systems can only be realized with privacy conserving personalization. In this paper, we propose an on-device visual question answering system that generates personalized answers using on-device user knowledge graph. These systems have the potential to serve as a fundamental groundwork for the development of genuinely intelligent and tailored assistants, targeted specifically to the needs and preferences of each individual. We validate our model performance on both in-realm, public datasets and personal user data. Our results show consistent performance increase across both tasks, with an absolute improvement of 36% with KVQA data-set on 1-hop inferences and 6% improvement on user personal data. We also conduct and showcase user-study results to validate our hypothesis of the need and relevance of proposed system.

****

Complementary-Contradictory Feature Regularization Against Multimodal Overfitting

Antonio Tejero-de-Pablos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5679-5688

Understanding multimodal learning is essential to design intelligent systems that can effectively combine various data types (visual, audio, etc.). Multimodal learning is not trivial, as adding new modalities does not always result in a significant improvement in performance, i.e., multimodal overfitting. To tackle this, several works proposed regularizing each modality's learning speed and feature distribution. However, in these methods, characterizing quantitatively and qualitatively multimodal overfitting is not intuitive. We hypothesize that, rather than regularizing abstract hyperparameters, regularizing the features learned is a more straightforward methodology against multimodal overfitting. For the given input modalities and task, we constrain "complementary" (useful) and "contradictory" (obstacle) features via a masking operation on the multimodal latent space. In addition, we leverage latent discretization so the size of the complementary-contradictory spaces becomes learnable, allowing the estimation of a modal complementarity measure. Our method successfully improves the performance of datasets with modality overfitting in different tasks, providing insight into "what" and "how much" is learned from each modality. Furthermore, it facilitates transfer learning to new datasets. Our code and a detailed manual are available at https://github.com/CyberAgentAILab/CM-VQVAE.

****

Appearance-Based Curriculum for Semi-Supervised Learning With Multi-Angle Unlabeled Data

Yuki Tanaka, Shuhei M. Yoshida, Takashi Shibata, Makoto Terao, Takayuki Okatani, Masashi Sugiyama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2780-2789

We propose an appearance-based curriculum (ABC) for a semi-supervised learning scenario where labeled images taken from limited angles and unlabeled ones taken from various angles are available for training. A common approach to semi-supervised learning relies on pseudo-labeling and data augmentation, but it struggles with large visual variations that cannot be covered by data augmentation. To solve this problem, ABC incrementally expands the pool of unlabeled images fed to a base semi-supervised learner so that newly added data are the ones most similar to those already in the pool. This way, the learner can assign pseudo-labels to the new data with high accuracy, keeping the quality of pseudo-labels higher than that when all the unlabeled data are processed at once, as customarily done in existing semi-supervised learning methods. We conducted extensive experiments and confirmed that our method outperforms the state-of-the-art semi-supervised learning methods in our scenario.

****

Incorporating Physics Principles for Precise Human Motion Prediction

Yufei Zhang, Jeffrey O. Kephart, Qiang Ji; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6164-6174

A variety of real-world applications rely on accurate predictions of 3D human motion from their past observations. While existing methods have made notable progress, their predictions over subsecond horizons can still be off by many centimeters. In this paper, we argue that achieving precise human motion prediction requires characterizing the fundamental physics principles governing body movements. We introduce PhysMoP, a novel framework that incorporates Physics for human Motion Prediction. PhysMoP estimates the body configuration of the next frame by solving the Euler-Lagrange equations, a set of Ordinary Different Equations describing the physical motion rules. To limit the inherent problem of error accumulation over time, PhysMoP leverages a data-driven model and iteratively guides the physics-based prediction via a fusion model. Through extensive experiments, we demonstrate that PhysMoP significantly outperforms existing approaches at subsecond prediction horizons. For example, at a prediction horizon of 80 msec, PhysMoP outperforms traditional data-driven approaches by a factor of 10 or more.

****

MuSHRoom: Multi-Sensor Hybrid Room Dataset for Joint 3D Reconstruction and Novel View Synthesis

Xuqian Ren, Wenjia Wang, Dingding Cai, Tuuli Tuominen, Juho Kannala, Esa Rahtu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4508-4517

Metaverse technologies demand accurate, real-time, and immersive modeling on consumer-grade hardware for both non-human perception (e.g., drone/robot/autonomous car navigation) and immersive technologies like AR/VR, requiring both structural accuracy and photorealism. However, there exists a knowledge gap in how to apply geometric reconstruction and photorealism modeling (novel view synthesis) in a unified framework. To address this gap and promote the development of robust and immersive modeling and rendering with consumer-grade devices, first, we propose a real-world Multi-Sensor Hybrid Room Dataset (MuSHRoom). Our dataset presents exciting challenges and requires state-of-the-art methods to be cost-effective, robust to noisy data and devices, and can jointly learn 3D reconstruction and novel view synthesis, instead of treating them as separate tasks, making them ideal for real-world applications. Second, we benchmark several famous pipelines on our dataset for joint 3D mesh reconstruction and novel view synthesis. Finally, in order to further improve the overall performance, we propose a new method that achieves a good trade-off between the two tasks. Our dataset and benchmark show great potential in promoting the improvements for fusing 3D reconstruction and high-quality rendering in a robust and computationally efficient end-to-end fashion. The dataset and code is available at the project webpate: https://xuqianren.github. io/publications/MuSHRoom/.

****

POISE: Pose Guided Human Silhouette Extraction Under Occlusions

Arindam Dutta, Rohit Lal, Dripta S. Raychaudhuri, Calvin-Khang Ta, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6153-6163

Human silhouette extraction is a fundamental task in computer vision with applications in various downstream tasks. However, occlusions pose a significant challenge, leading to distorted silhouettes. To address this challenge, we introduce POISE : Pose Guided Human Silhouette Extraction under Occlusions, a fusion framework that enhances accuracy and robustness in human silhouette prediction. By combining initial silhouette estimates from a segmentation model with human joint predictions from a 2D pose estimation model, POISE leverages the complementary strengths of both approaches, effectively integrating precise body shape information and spatial information to tackle occlusions. Furthermore, the unsupervised nature of POISE eliminates the need for costly annotations, making it scalable and practical. Extensive experimental results demonstrate its superiority in improving silhouette extraction under occlusions, with promising results in downstream tasks such as gait recognition.

****

Shape-Guided Diffusion With Inside-Outside Attention

Dong Huk Park, Grace Luo, Clayton Toste, Samaneh Azadi, Xihui Liu, Maka Karalashvili, Anna Rohrbach, Trevor Darrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4198-4207

We introduce precise object silhouette as a new form of user control in text-to-image diffusion models, which we dub Shape-Guided Diffusion. Our training-free method uses an Inside-Outside Attention mechanism during the inversion and generation process to apply a shape constraint to the cross- and self-attention maps. Our mechanism designates which spatial region is the object (inside) vs. background (outside) then associates edits to the correct region. We demonstrate the efficacy of our method on the shape-guided editing task, where the model must replace an object according to a text prompt and object mask. We curate a new ShapePrompts benchmark derived from MS-COCO and achieve SOTA results in shape faithfulness without a degradation in text alignment or image realism according to both automatic metrics and annotator ratings. Our data and code will be made available at https://shape-guided-diffusion.github.io.

****

Learning Visual Body-Shape-Aware Embeddings for Fashion Compatibility

Kaicheng Pang, Xingxing Zou, Waikeung Wong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8056-8065

Body shape is a crucial factor in outfit recommendation. Previous studies that directly used body measurement data to investigate the relationship between body shape and outfit have achieved limited performance due to oversimplified body shape representations. This paper proposes a Visual Body-shape-Aware Network (ViBA-Net) to improve the fashion compatibility model's awareness of human body shape through visual-level information. Specifically, ViBA-Net consists of three modules: a body-shape embedding module, which extracts visual and anthropometric features of body shape from a newly introduced large-scale body shape dataset; an outfit embedding module, which learns the outfit representation based on visual features extracted from a try-on image and textual features extracted from fashion attributes; and a joint embedding module, which jointly models the relationship between the representations of body shape and outfit. ViBA-Net is designed to generate attribute-level explanations for the evaluation results based on the computed attention weights. The effectiveness of ViBA-Net is evaluated on two mainstream datasets through qualitative and quantitative analysis. Data and code are released.

****

Unsupervised Exemplar-Based Image-to-Image Translation and Cascaded Vision Transformers for Tagged and Untagged Cardiac Cine MRI Registration

Meng Ye, Mikael Kanski, Dong Yang, Leon Axel, Dimitris Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7644-7654

Multi-modal registration between tagged and untagged cardiac cine magnetic resonance (MR) images remains difficult, due to the domain gap and large deformations between the two modalities. Recent work using an image-to-image translation (I2I) module to overcome the domain gap can convert the multi-modal into a mono-modal registration task and take advantage of advanced mono-modal registration architectures. However, they often ignore two issues: the sample-specific style of each image to be registered during I2I and large hybrid rigid and non-rigid deformations between modalities. We first propose an exemplar-based I2I module capable of unsupervised cross-domain correspondence learning to enforce the style consistency between the fake image and the image to be registered. Then we propose an efficient cascaded vision transformer-based registration network to predict both the affine and non-rigid deformations, in which a single feature embedding subnetwork is shared by the two stages of deformation prediction. We validated our method on a clinical cardiac MR dataset with paired but unaligned untagged and tagged MR images. The results show that our method outperforms traditional methods significantly in terms of the I2I quality and multi-modal image registration accuracy.

****

Spectroformer: Multi-Domain Query Cascaded Transformer Network for Underwater Image Enhancement
Raqib Khan, Priyanka Mishra, Nancy Mehta, Shruti S. Phutke, Santosh Kumar Vipparthi, Sukumar Nandi, Subrahmanyam Murala; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1454-1463
Underwater images often suffer from color distortion, haze, and limited visibility due to light refraction and absorption in water. These challenges significantly impact autonomous underwater vehicle applications, necessitating efficient image enhancement techniques. To address these challenges, we propose a Multi-Domain Query Cascaded Transformer Network for underwater image enhancement. Our approach includes a novel Multi-Domain Query Cascaded Attention mechanism that integrates localized transmission features and global illumination features. To improve feature propagation from the encoder to the decoder, we propose a Spatio-Spectro Fusion-Based Attention Block. Additionally, we introduce a Hybrid Fourier-Spatial Upsampling Block, which uniquely combines Fourier and spatial upsampling techniques to enhance feature resolution effectively. We evaluate our method on benchmark synthetic and real-world underwater image datasets, demonstrating its superiority through extensive ablation studies and comparative analysis. The testing code is available at: https: //github.com/Mdraqibkhan/Spectroformer.
****

Removing the Quality Tax in Controllable Face Generation
Yiwen Huang, Zhiqiu Yu, Xinjie Yi, Yue Wang, James Tompkin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5364-5373
3DMM conditioned face generation has gained traction due to its well-defined controllability; however, the trade-off is lower sample quality: Previous works such as DiscoFaceGAN and 3D-FM GAN show a significant FID gap compared to the unconditional StyleGAN, suggesting that there is a quality tax to pay for controllability. In this paper, we challenge the assumption that quality and controllability cannot coexist. To pinpoint the previous issues, we mathematically formalize the problem of 3DMM conditioned face generation. Then, we devise simple solutions to the problem under our proposed framework. This results in a new model that effectively removes the quality tax between 3DMM conditioned face GANs and the unconditional StyleGAN. Project webpage: https://visual.cs.brown.edu/taxfreegan
****

On Manipulating Scene Text in the Wild With Diffusion Models
Joshua Santoso, Christian Simon, Williem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5202-5211
Diffusion models have gained attention for image editing yielding impressive results in text-to-image tasks. On the downside, one might notice that generated images of stable diffusion models suffer from deteriorated details. This pitfall impacts image editing tasks that require information preservation e.g., scene text editing. As a desired result, the model must show the capability to replace the text on the source image to the target text while preserving the details e.g., color, font size, and background. To leverage the potential of diffusion models, in this work, we introduce Diffusion-BasEd Scene Text manipulation network so-called DBEST. Specifically, we design two adaptation strategies, namely one-shot style adaptation and text-recognition guidance. In experiments, we thoroughly assess and compare our proposed method against state-of-the-arts on various scene text datasets, then provide extensive ablation studies for each granularity to analyze our performance gain. Also, we demonstrate the effectiveness of our proposed method to synthesize scene text indicated by competitive Optical Character Recognition (OCR) accuracy. Our method achieves 94.15% and 98.12% on COCO-text and ICDAR2013 datasets for character-level evaluation.
****

Improved Techniques for Quantizing Deep Networks With Adaptive Bit-Widths
Ximeng Sun, Rameswar Panda, Chun-Fu Richard Chen, Naigang Wang, Bowen Pan, Aude Oliva, Rogerio Feris, Kate Saenko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 957-967
Quantizing deep networks with adaptive bit-widths is a promising technique for e

fficient inference across many devices and resource constraints. In contrast to static methods that repeat the quantization process and train different models for different constraints, adaptive quantization enables us to flexibly adjust the bit-widths of a single deep network during inference for instant adaptation in different scenarios. While existing research shows encouraging results on common image classification benchmarks, this paper investigates how to train such adaptive networks more effectively. Specifically, we present two novel techniques for quantizing deep neural networks with adaptive bit-widths of weights and activations. First, we propose a collaborative strategy to choose a high-precision "teacher" for transferring knowledge to the low-precision "student" while jointly optimizing the model with all bit-widths. Second, to effectively transfer knowledge, we develop a dynamic block swapping method by randomly replacing the blocks in the lower-precision student network with the corresponding blocks in the higher-precision teacher network. Extensive experiments on multiple image classification datasets and novel video classification experiments, well demonstrate the efficacy of our approach over state-of-the-art methods.
****

Mining and Unifying Heterogeneous Contrastive Relations for Weakly-Supervised Actor-Action Segmentation
Bin Duan, Hao Tang, Changchang Sun, Ye Zhu, Yan Yan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 494-503
We introduce a novel weakly-supervised video actor-action segmentation (VAAS) framework, where only video-level tags are available. Previous VAAS methods follow a synthesize-and-refine scheme, i.e., they first synthesize the pseudo-segmentation and recursively refine the segmentation. However, this process requires significant time costs and heavily relies on the quality of the initial segmentation. Unlike existing works, our method hierarchically mines contrastive relations to supplement each other for learning a visually-plausible segmentation model. Specifically, three contrastive relations are abstracted from the pixel-level and frame-level, i.e., low-level edge-aware, class-activation map aware, and semantic tag-aware relations. Then, the discovered contrastive relations are unified into a universal objective for training the segmentation model, regardless of their heterogeneity. Moreover, we incorporate motion cues and unlabeled samples to increase the discriminative power and robustness of the segmentation model. Extensive experiments indicate that our proposed method produces reasonable segmentation.
****

Rethinking Knowledge Distillation With Raw Features for Semantic Segmentation
Tao Liu, Chenshu Chen, Xi Yang, Wenming Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1155-1164
Most existing knowledge distillation methods for semantic segmentation focus on extracting various sophisticated knowledge from raw features. However, such knowledge is usually manually designed and relies on prior knowledge as in traditional feature engineering. In this paper, we aim to propose a simple and effective feature distillation method using raw features. To this end, we revisit the pioneering work in feature distillation, FitNets, which simply minimizes the mean squared error (MSE) loss between the teacher and student features. Our experiments show that this naive method yields good results, even surpassing some well-designed methods in some cases. However, it requires carefully tuning the weight of distillation loss. By decomposing the loss function of FitNets into a magnitude difference term and an angular difference term, we find the weight of the angular difference term is affected by the magnitudes of the teacher features and the student features. We experimentally show that the angular difference term plays a crucial role in feature distillation and the magnitude of the features produced by different models may vary significantly. Therefore, it is hard to determine a suitable loss weight for various models. To avoid the weight of the angular distillation term being affected by the magnitude of the features, we propose Angular Distillation and explore distilling angular information along different feature dimensions for semantic segmentation. Extensive experiments show that our simple method exhibits great robustness to hyper-parameters and achieves state-of

-the-art distillation performance for semantic segmentation.
****

Fully-Automatic Reflection Removal for 360-Degree Images
Jonghyuk Park, Hyeona Kim, Eunpil Park, Jae-Young Sim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1609-1617
Reflection removal (RR) is a technique to reconstruct the transmitted scene behind the glass from a mixed image taken through glass. In 360-degree images, the mixed image region and the reference image region capturing the reflected scene exist together, and the mixed image is often restored by using the information of reference image. In this paper, we first propose a fully-automatic end-to-end RR framework for 360-degree images which automatically detects the mixed and reference image regions and removes the reflection artifacts in the mixed image by using the reference information simultaneously. We devise a transformer based U-Net architecture with horizontal windowing scheme to capture the long-range dependencies between the mixed and reference images via the self-attention mechanism and suppress the reflection artifacts by using the reference information. We also construct a training dataset of 360-degree images by synthesizing realistic reflection artifacts considering diverse geometric relation and photometric variation between the mixed and reference images. The experimental results show that the proposed method detects the mixed and reference image regions reliably without user-annotation and achieves better performance of RR compared with the state-of-the-art methods.
****

MITFAS: Mutual Information Based Temporal Feature Alignment and Sampling for Aerial Video Action Recognition
Ruiqi Xian, Xijun Wang, Dinesh Manocha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6625-6634
We present a novel approach for action recognition in UAV videos. Our formulation is designed to handle occlusion and viewpoint changes caused by the movement of a UAV. We use the concept of mutual information to compute and align the regions corresponding to human action or motion in the temporal domain. This enables our recognition model to learn from the key features associated with the motion. We also propose a novel frame sampling method that uses joint mutual information to acquire the most informative frame sequence in UAV videos. We have integrated our approach with X3D and evaluated the performance on multiple datasets. In practice, we achieve 18.9% improvement in Top-1 accuracy over current state-of-the-art methods on UAV-Human, 7.3% improvement on Drone-Action, and 7.16% improvement on NEC Drones. The code is available at https://github.com/Ricky-Xian/MITFAS.
****

Multimodal Deep Learning for Remote Stress Estimation Using CCT-LSTM
Sayyedjavad Ziaratnia, Tipporn Laohakangvalvit, Midori Sugaya, Peeraya Sripian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8336-8344
Stress estimation is key to the early detection and mitigation of health problems, enhancing driving safety through driver stress monitoring, and improving human-robot interaction efficiency by adapting to user's stress levels. In this paper, we present a novel method for video-based remote stress estimation and categorization, which involves two separate experiments: one for stress task classification and another for multilevel stress classification. The method combines two deep learning approaches, the Compact Convolutional Transformer (CCT) and Long Short-Term Memory (LSTM), to form a CCT-LSTM pipeline. For each modality (facial expression and rPPG), a CCT model is used to extract features, followed by an LSTM block for temporal pattern recognition. In stress task classification, T1, T2, and T3 tasks from the UBFC-Phys dataset are used, utilizing sevenfold cross-validation. The results indicated a mean accuracy of 83.2% and an F1 score of 83.4%. For multilevel stress classification, the control (lower stress) and test (higher stress) groups from the same dataset were used with fivefold cross-validation, achieving a mean accuracy of 80.5% and an F1 score of 80.3%. The results sug

gest that our proposed model surpasses existing stress estimation methods by effectively using multimodal deep learning and the CCT-LSTM pipeline for precise, non-invasive stress detection and categorization, with applications in health monitoring, safety, and interactive technologies.

****

Let the Beat Follow You - Creating Interactive Drum Sounds From Body Rhythm

Xiulong Liu, Kun Su, Eli Shlizerman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7187-7197

It is often the case that human body movements include rhythmic patterns. A video camera system that captures these patterns and responds to them with rhythmic sounds or music as these happen could create a unique interactive experience. Creating such an experience is challenging and cannot be achieved with existing methods since it requires a real-time translation of related visual cues into in-rhythm sounds. In this work, we propose a novel learning-based system, called 'InteractiveBeat', which generates an evolving interactive soundtrack for a camera input that captures person's movements. InteractiveBeat infers body skeleton keypoints and translates them into drum rhythms using a series of sequence models. It then implements a conditional drum generation network for generating polyphonic drum sounds based on the rhythms. To guarantee real-time function, these models are integrated into a time-evolving pipeline with update rules. For training and evaluation of InteractiveBeat, in addition to training on well-annotated large-scale dance database, we collected a dataset of in-the-wild videos with people performing movements of various activities that correspond to background music. We evaluate InteractiveBeat in two scenarios: i) laboratory setting, ii) prerecorded videos of movements from in-the-wild videos, and develop 'live' demo prototype of the system. Our results on evaluations show that the system can generate interactive rhythmic drums with higher accuracy than existing methods and achieves non-cumulative latency of 34ms. This allows InteractiveBeat to be synchronized with the video stream and to react to movements in real-time.

****

A Visual Active Search Framework for Geospatial Exploration

Anindya Sarkar, Michael Lanier, Scott Alfeld, Jiarui Feng, Roman Garnett, Nathan Jacobs, Yevgeniy Vorobeychik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8316-8325

Many problems can be viewed as forms of geospatial search aided by aerial imagery, with examples ranging from detecting poaching activity to human trafficking. We model this class of problems in a visual active search (VAS) framework, which has three key inputs: (1) an image of the entire search area, which is subdivided into regions, (2) a local search function, which determines whether a previously unseen object class is present in a given region, and (3) a fixed search budget, which limits the number of times the local search function can be evaluated. The goal is to maximize the number of objects found within the search budget. We propose a reinforcement learning approach for VAS that learns a meta-search policy from a collection of fully annotated search tasks. This meta-search policy is then used to dynamically search for a novel target-object class, leveraging the outcome of any previous queries to determine where to query next. Through extensive experiments on several large-scale satellite imagery datasets, we show that the proposed approach significantly outperforms several strong baselines. We also propose novel domain adaptation techniques that improve the policy at decision time when there is a significant domain gap with the training data. Code is publicly available.

****

ShARc: Shape and Appearance Recognition for Person Identification In-the-Wild

Haidong Zhu, Wanrong Zheng, Zhaoheng Zheng, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6290-6300

Identifying individuals in unconstrained video settings is a valuable yet challenging task in biometric analysis due to variations in appearances, environments, degradations, and occlusions. In this paper, we present ShARc, a multimodal approach for video-based person identification in uncontrolled environments that em phasizes 3-D body shape, pose, and appearance. We introduce two encoders: a Pose and Shape Encoder (PSE) and an Aggregated Appearance Encoder (AAE). PSE encodes the body shape via binarized silhouettes, skeleton motions, and 3-D body shape, while AAE provides two levels of temporal appearance feature aggregation: attention-based feature aggregation and averaging aggregation. For attention-based feature aggregation, we employ spatial and temporal attention to focus on key areas for person distinction. For averaging aggregation, we introduce a novel flattening layer after averaging to extract more distinguishable information and reduce overfitting of attention. We utilize centroid feature averaging for gallery registration. We demonstrate significant improvements over existing state-of-the-art methods on public datasets, including CCVID, MEVID, and BRIAR.

****

DocReal: Robust Document Dewarping of Real-Life Images via Attention-Enhanced Control Point Prediction

Fangchen Yu, Yina Xie, Lei Wu, Yafei Wen, Guozhi Wang, Shuai Ren, Xiaoxin Chen, Jianfeng Mao, Wenye Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 665-674

Document image dewarping is a crucial task in computer vision with numerous practical applications. The control point method, as a popular image dewarping approach, has attracted attention due to its simplicity and efficiency. However, inaccurate control point prediction due to varying background noises and deformation types can result in unsatisfactory performance. To address these issues, we propose a robust document dewarping approach for real-life images, namely DocReal, which utilizes Enet to effectively remove background noise and an attention-enhanced control point (AECP) module to better capture local deformations. Moreover, we augment the training data by synthesizing 2D images with 3D deformations and additional deformation types. Our proposed method achieves state-of-the-art performance on the DocUNet benchmark and a newly proposed benchmark of 200 Chinese distorted images, exhibiting superior dewarping accuracy, OCR performance, and robustness to various types of image distortion.

****

Multi-Level Attention Aggregation for Aesthetic Face Relighting

Hemanth Pidaparthy, Abhay Chauhan, Pavan Sudheendra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4057-4066

Face relighting is the challenging task of estimating the illumination cast on portrait images by a light source varying in both position and intensity. As shadows are an important aspect of relighting, many prior works focus on estimating accurate shadows using either a shadow mask or face geometry. While these work well, the rendered images do not look aesthetic/photo-realistic. We propose a novel method that combines the features from attention maps at higher resolutions with the lighting information to estimate aesthetic relit images with accurate shadows. We created a new relighting dataset using a synthetic One-Light-At-a-Time (OLAT) lighting rig in Blender software that captures most of the variations encountered in face relighting. Through extensive experimental validation, we show that the performance of our model is better than the current state-of-art face relighting models despite training on a significantly smaller dataset of only synthetic images. We also demonstrate unsupervised domain adaptation from synthetic to real images. We show that our model is able to adapt very well to significantly different out-of-training light source positions.

****

Learning Residual Elastic Warps for Image Stitching Under Dirichlet Boundary Condition

Minsu Kim, Yongjun Lee, Woo Kyoung Han, Kyong Hwan Jin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4016-4024

Trendy suggestions for learning-based elastic warps enable the deep image stitchings to align images exposed to large parallax errors. Despite the remarkable alignments, the methods struggle with occasional holes or discontinuity between overlapping and non-overlapping regions of a target image as the applied training

strategy mostly focuses on overlap region alignment. As a result, they require additional modules such as seam finder and image inpainting for hiding discontinuity and filling holes, respectively. In this work, we suggest Recurrent Elastic Warps (REwarp) that address the problem with Dirichlet boundary condition and boost performances by residual learning for recurrent misalign correction. Specifically, REwarp predicts a homography and a Thin-plate Spline (TPS) under the boundary constraint for discontinuity and hole-free image stitching. Our experiments show the favorable aligns and the competitive computational costs of REwarp compared to the existing stitching methods. Our source code is available at https://github.com/minshu-kim/REwarp.
****
Interactive Segmentation for Diverse Gesture Types Without Context
Josh Myers-Dean, Yifei Fan, Brian Price, Wilson Chan, Danna Gurari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7198-7208
Interactive segmentation entails a human marking an image to guide how a model either creates or edits a segmentation. Our work addresses limitations of existing methods: they either only support one gesture type for marking an image (e.g., either clicks or scribbles) or require knowledge of the gesture type being employed, and require specifying whether marked regions should be included versus excluded in the final segmentation. We instead propose a simplified interactive segmentation task where a user only must mark an image, where the input can be of any gesture type without specifying the gesture type. We support this new task by introducing the first interactive segmentation dataset with multiple gesture types as well as a new evaluation metric capable of holistically evaluating interactive segmentation algorithms. We then analyze numerous interactive segmentation algorithms, including ones adapted for our novel task. While we observe promising performance overall, we also highlight areas for future improvement. To facilitate further extensions of this work, we publicly share our new dataset at https://github.com/joshmyersdean/dig.
****
Customizing 360-Degree Panoramas Through Text-to-Image Diffusion Models
Hai Wang, Xiaoyu Xiang, Yuchen Fan, Jing-Hao Xue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4933-4943
Personalized text-to-image (T2I) synthesis based on diffusion models has attracted significant attention in recent research. However, existing methods primarily concentrate on customizing subjects or styles, neglecting the exploration of global geometry. In this study, we propose an approach that focuses on the customization of 360-degree panoramas, which inherently possess global geometric properties, using a T2I diffusion model. To achieve this, we curate a paired image-text dataset specifically designed for the task and subsequently employ it to fine-tune a pre-trained T2I diffusion model with LoRA. Nevertheless, the fine-tuned model alone does not ensure the continuity between the leftmost and rightmost sides of the synthesized images, a crucial characteristic of 360-degree panoramas. To address this issue, we propose a method called StitchDiffusion. Specifically, we perform pre-denoising operations twice at each time step of the denoising process on the stitch block consisting of the leftmost and rightmost image regions. Furthermore, a global cropping is adopted to synthesize seamless 360-degree panoramas. Experimental results demonstrate the effectiveness of our customized model combined with the proposed StitchDiffusion in generating high-quality 360-degree panoramic images. Moreover, our customized model exhibits exceptional generalization ability in producing scenes unseen in the fine-tuning dataset. Code is available at https://github.com/littlewhitesea/StitchDiffusion.
****
Temporal Context Enhanced Referring Video Object Segmentation
Xiao Hu, Basavaraj Hampiholi, Heiko Neumann, Jochen Lang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5574-5583
The goal of Referring Video Object Segmentation is to extract an object from a video clip based on a given expression. While previous methods have utilized the

transformer's multi-modal learning capabilities to aggregate information from different modalities, they have mainly focused on spatial information and paid less attention to temporal information. To enhance the learning of temporal information, we propose TCE-RVOS with a novel frame token fusion (FTF) structure and a novel instance query transformer (IQT). Our technical innovations maximize the potential information gain of videos over single images. Our contributions also include a new classification of two widely used validation datasets for investigation of challenging cases. Our experimental results demonstrate that TCE-RVOS effectively captures temporal information and outperforms the previous state-of-the-art methods by increasing the J&F score by 4.0 and 1.9 points using ResNet-50 and VSwin-Tiny as the backbone on Ref-Youtube-VOS, respectively, and +2.0 mAP on A2D-Sentences dataset by using VSwin-Tiny backbone. The code is available at https://github.com/haliphinx/TCE-RVOS

****

Revisiting Token Pruning for Object Detection and Instance Segmentation

Yifei Liu, Mathias Gehrig, Nico Messikommer, Marco Cannici, Davide Scaramuzza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2658-2668

Vision Transformers (ViTs) have shown impressive performance in computer vision, but their high computational cost, quadratic in the number of tokens, limits their adoption in computation-constrained applications. However, this large number of tokens may not be necessary, as not all tokens are equally important. In this paper, we investigate token pruning to accelerate inference for object detection and instance segmentation, extending prior works from image classification. Through extensive experiments, we offer four insights for dense tasks: (i) tokens should not be completely pruned and discarded, but rather preserved in the feature maps for later use. (ii) reactivating previously pruned tokens can further enhance model performance. (iii) a dynamic pruning rate based on images is better than a fixed pruning rate. (iv) a lightweight, 2-layer MLP can effectively prune tokens, achieving accuracy comparable with complex gating networks with a simpler design. We assess the effects of these design decisions on the COCO dataset and introduce an approach that incorporates these findings, showing a reduction in performance decline from 1.5 mAP to 0.3 mAP in both boxes and masks, compared to existing token pruning methods. In relation to the dense counterpart that utilizes all tokens, our method realizes an increase in inference speed, achieving up to 34% faster performance for the entire network and 46% for the backbone. Code will be publicly available.

****

AssemblyNet: A Point Cloud Dataset and Benchmark for Predicting Part Directions in an Exploded Layout

Jesper Gaarsdal, Joakim Bruslund Haurum, Sune Wolff, Claus Brøndgaard Madsen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5836-5845

Exploded views are powerful tools for visualizing the assembly and disassembly of complex objects, widely used in technical illustrations, assembly instructions, and product presentations. Previous methods for automating the creation of exploded views are either slow and computationally costly or compromise on accuracy. Therefore, the construction of exploded views is typically a manual process. In this paper, we propose a novel approach for automatically predicting the direction of parts in an exploded view using deep learning. To achieve this, we introduce a new dataset, AssemblyNet, which contains point cloud data sampled from 3D models of real-world assemblies, including water pumps, mixed industrial assemblies, and LEGO models. The AssemblyNet dataset includes a total of 44 assemblies, separated into 495 subassemblies with a total of 5420 parts. We provide ground truth labels for regression and classification, representing the directions in which the parts are moved in the exploded views. We also provide performance benchmarks using various state-of-the-art models for shape classification on point clouds and propose a novel two-path network architecture. Project page available at https://github.com/jgaarsdal/AssemblyNet

****

Location-Aware Self-Supervised Transformers for Semantic Segmentation

Mathilde Caron, Neil Houlsby, Cordelia Schmid; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 117-127

Pixel-level labels are particularly expensive to acquire. Hence, pretraining is a critical step to improve models on a task like semantic segmentation. However, prominent algorithms for pretraining neural networks use image-level objectives, e.g. image classification, image-text alignment a la CLIP, or self-supervised contrastive learning. These objectives do not model spatial information, which might be sub-optimal when finetuning on downstream tasks with spatial reasoning. In this work, we pretrain networks with a location-aware (LOCA) self-supervised method which fosters the emergence of strong dense features. Specifically, we use both a patch-level clustering scheme to mine dense pseudo-labels and a relative location prediction task to encourage learning about object parts and their spatial arrangement. Our experiments show that LOCA pretraining leads to representations that transfer competitively to challenging and diverse semantic segmentation datasets.

****

Self-Supervised Learning for Visual Relationship Detection Through Masked Bounding Box Reconstruction

Zacharias Anastasakis, Dimitrios Mallis, Markos Diomataris, George Alexandridis, Stefanos Kollias, Vassilis Pitsikalis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1206-1215

We present a novel self-supervised approach for representation learning, particularly for the task of Visual Relationship Detection (VRD). Motivated by the effectiveness of Masked Image Modeling (MIM), we propose Masked Bounding Box Reconstruction (MBBR), a variation of MIM where a percentage of the entities/objects within a scene are masked and subsequently reconstructed based on the unmasked objects. The core idea is that, through object-level masked modeling, the network learns context-aware representations that capture the interaction of objects within a scene and thus are highly predictive of visual object relationships. We extensively evaluate learned representations, both qualitatively and quantitatively, in a few-shot setting and demonstrate the efficacy of MBBR for learning robust visual representations, particularly tailored for VRD. The proposed method is able to surpass state-of-the-art VRD methods on the Predicate Detection (PredDet) evaluation setting, using only a few annotated samples. We make our code available at https://github.com/deeplab-ai/SelfSupervisedVRD.

****

Real-Time 6-DoF Pose Estimation by an Event-Based Camera Using Active LED Markers

Gerald Ebmer, Adam Loch, Minh Nhat Vu, Roberto Mecca, Germain Haessig, Christian Hartl-Nesic, Markus Vincze, Andreas Kugi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8137-8146

Real-time applications for autonomous operations depend largely on fast and robust vision-based localization systems. Since image processing tasks require processing large amounts of data, the computational resources often limit the performance of other processes. To overcome this limitation, traditional marker-based localization systems are widely used since they are easy to integrate and achieve reliable accuracy. However, classical marker-based localization systems significantly depend on standard cameras with low frame rates, which often lack accuracy due to motion blur. In contrast, event-based cameras provide high temporal resolution and a high dynamic range, which can be utilized for fast localization tasks, even under challenging visual conditions. This paper proposes a simple but effective event-based pose estimation system using active LED markers (ALM) for fast and accurate pose estimation. The proposed algorithm is able to operate in real time with a latency below 0.5 ms while maintaining output rates of 3 kHz. Experimental results in static and dynamic scenarios are presented to demonstrate the performance of the proposed approach in terms of computational speed and absolute accuracy, using the OptiTrack system as the basis for measurement. Moreover, we demonstrate the feasibility of the proposed approach by deploying the hardware, i.e., the event-based camera and ALM, and the software in a real quadcopt er application.

****

P-Age: Pexels Dataset for Robust Spatio-Temporal Apparent Age Classification

Abid Ali, Ashish Marisetty, François Brémond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8606-8615

Age estimation is a challenging task that has numerous applications. In this paper, we propose a new direction for age classification that utilizes a video-based model to address challenges such as occlusions, low-resolution, and lighting conditions. To address these challenges, we propose AgeFormer which utilizes spatio-temporal information on the dynamics of the entire body dominating face-based methods for age classification. Our novel two-stream architecture uses TimeSformer and EfficientNet as backbones, to effectively capture both facial and body dynamics information for efficient and accurate age estimation in videos. Furthermore, to fill the gap in predicting age in real-world situations from videos, we construct a video dataset called Pexels Age (P-Age) for age classification. The proposed method achieves superior results compared to existing face-based age estimation methods and is evaluated in situations where the face is highly occluded, blurred, or masked. The method is also cross-tested on a variety of challenging video datasets such as Charades, Smarthome, and Thumos-14.

****

SSVOD: Semi-Supervised Video Object Detection With Sparse Annotations

Tanvir Mahmud, Chun-Hao Liu, Burhaneddin Yaman, Diana Marculescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6773-6782

Despite significant progress in semi-supervised learning for image object detection, several key issues are yet to be addressed for video object detection: (1) Achieving good performance for supervised video object detection greatly depends on the availability of annotated frames. (2) Despite having large inter-frame correlations in a video, collecting annotations for a large number of frames per video is expensive, time-consuming, and often redundant. (3) Existing semi-supervised techniques on static images can hardly exploit the temporal motion dynamics inherently present in videos. In this paper, we introduce SSVOD, an end-to-end semi-supervised video object detection framework that exploits motion dynamics of videos to utilize large-scale unlabeled frames with sparse annotations. To selectively assemble robust pseudo-labels across groups of frames, we introduce flow-warped predictions from nearby frames for temporal-consistency estimation. In particular, we introduce cross-IoU and cross-divergence based selection methods over a set of estimated predictions to include robust pseudo-labels for bounding boxes and class labels, respectively. To strike a balance between confirmation bias and uncertainty noise in pseudo-labels, we propose confidence threshold based combination of hard and soft pseudo-labels. Our method achieves significant performance improvements over existing methods on ImageNet-VID, Epic-KITCHENS, and YouTube-VIS datasets. Codes are available at https://github.com/enyac-group/SSVOD.git.

****

Deep Optics for Optomechanical Control Policy Design

Justin Fletcher; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8306-8315

An emerging class of Fizeau optical telescopes have the potential to upend prior cost scaling models, substantially improving the angular resolution and contrast attainable by ground-based astronomical instruments. However, this design introduces a challenging visual control problem that must be solved to compensate for wavefront aberrations induced by the flexible substructure it employs. We subvert this problem with a deep optics approach to policy design and image recovery that exploits, rather than corrects, aberrations to obtain domain-specific object recovery performance exceeding that of more costly filled aperture designs.

****

A Generative Multi-Resolution Pyramid and Normal-Conditioning 3D Cloth Draping

Hunor Laczkó, Meysam Madadi, Sergio Escalera, Jordi Gonzalez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp.

8709-8718

RGB cloth generation has been deeply studied in the related literature, however, 3D garment generation remains an open problem. In this paper, we build a conditional variational autoencoder for 3D garment generation and draping. We propose a pyramid network to add garment details progressively in a canonical space, i.e. unposing and unshaping the garments w.r.t. the body. We study conditioning the network on surface normal UV maps, as an intermediate representation, which is an easier problem to optimize than 3D coordinates. Our results on two public datasets, CLOTH3D and CAPE, show that our model is robust, controllable in terms of detail generation by the use of multi-resolution pyramids, and achieves state-of-the-art results that can highly generalize to unseen garments, poses, and shapes even when training with small amounts of data. The code can be found at: https://github.com/HunorLaczko/pyramid-drape

****

MAdVerse: A Hierarchical Dataset of Multi-Lingual Ads From Diverse Sources and Categories

Amruth Sagar, Rishabh Srivastava, Rakshitha R. T., Venkata Kesav Venna, Ravi Kiran Sarvadevabhatla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8087-8096

The convergence of computer vision and advertising has sparked substantial interest lately. Existing advertisement datasets often derive from subsets of established data with highly specialized annotations or feature diverse annotations without a cohesive taxonomy among ad images. Notably, no datasets encompass diverse advertisement styles or semantic grouping at various levels of granularity for a better understanding of ads. Our work addresses this gap by introducing MAdVerse, an extensive, multilingual compilation of more than 50,000 ads from the web, social media websites and e-newspapers. Advertisements are hierarchically grouped with uniform granularity into 11 categories, divided into 51 sub-categories, and 524 fine-grained brands at leaf level, each featuring ads in various languages. Furthermore, we provide comprehensive baseline classification results for various pertinent prediction tasks within the realm of advertising analysis. Specifically, these tasks include hierarchical ad classification, source classification, multilingual classification and inducing hierarchy in existing ad datasets. The dataset, code and models are available on the project page https://madverse24.github.io/

****

Identifying Label Errors in Object Detection Datasets by Loss Inspection

Marius Schubert, Tobias Riedlinger, Karsten Kahl, Daniel Kröll, Sebastian Schoenen, Siniša Šegvi■, Matthias Rottmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4582-4591

Labeling datasets for supervised object detection is a dull and time-consuming task. Errors can be easily introduced during annotation and overlooked during review, yielding inaccurate benchmarks and performance degradation of deep neural networks trained on noisy labels. In this work, we introduce a benchmark for label error detection methods on object detection datasets as well as a theoretically underpinned label error detection method and a number of baselines. We simulate four different types of randomly introduced label errors on train and test sets of well-labeled object detection datasets. For our label error detection method we assume a two-stage object detector to be given and consider the sum of both stages' classification and regression losses. The losses are computed with respect to the predictions and the noisy labels including simulated label errors, aiming at detecting the latter. We compare our method to four baselines: a naive one without deep learning, the object detector's score, the entropy of the classification softmax distribution and a probability margin based method from related work. We outperform all baselines and demonstrate that among the considered methods, ours is the only one that detects label errors of all four types efficiently, which we also derive theoretically. Furthermore, we detect real label errors a) on commonly used test datasets in object detection and b) on a proprietary dataset. In both cases we achieve low false positives rates, i.e., we detect label errors with a precision for a) of up to 71.5% and for b) with 97%.

****

Reference-Based Restoration of Digitized Analog Videotapes

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, Alberto Del Bimbo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1659-1668

Analog magnetic tapes have been the main video data storage device for several decades. Videos stored on analog videotapes exhibit unique degradation patterns caused by tape aging and reader device malfunctioning that are different from those observed in film and digital video restoration tasks. In this work, we present a reference-based approach for the resToration of digitized Analog videotaPEs (TAPE). We leverage CLIP for zero-shot artifact detection to identify the cleanest frames of each video through textual prompts describing different artifacts. Then, we select the clean frames most similar to the input ones and employ them as references. We design a transformer-based Swin-UNet network that exploits both neighboring and reference frames via our Multi-Reference Spatial Feature Fusion (MRSFF) blocks. MRSFF blocks rely on cross-attention and attention pooling to take advantage of the most useful parts of each reference frame. To address the absence of ground truth in real-world videos, we create a synthetic dataset of videos exhibiting artifacts that closely resemble those commonly found in analog videotapes. Both quantitative and qualitative experiments show the effectiveness of our approach compared to other state-of-the-art methods. The code, the model, and the synthetic dataset are publicly available at https://github.com/miccunifi/TAPE.

****

BigSmall: Efficient Multi-Task Learning for Disparate Spatial and Temporal Physiological Measurements

Girish Narayanswamy, Yujia Liu, Yuzhe Yang, Chengqian Ma, Xin Liu, Daniel McDuff, Shwetak Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7914-7924

Understanding of human visual perception has historically inspired the design of computer vision architectures. As an example, perception occurs at different scales both spatially and temporally, suggesting that the extraction of salient visual information may be made more effective by attending to specific features at varying scales. Visual changes in the body due to physiological processes also occur at varying scales and with modality-specific characteristic properties. Inspired by this, we present BigSmall, an efficient architecture for physiological and behavioral measurement. We present the first joint camera-based facial action, cardiac, and pulmonary measurement model. We propose a multi-branch network with wrapping temporal shift modules that yields both accuracy and efficiency gains. We observe that fusing low-level features leads to suboptimal performance, but that fusing high level features enables efficiency gains with negligible losses in accuracy. Experimental results demonstrate that BigSmall significantly reduces the computational costs. Furthermore, compared to existing task-specific models, BigSmall achieves comparable or better results on multiple physiological measurement tasks simultaneously with a unified model.

****

Robust Feature Learning and Global Variance-Driven Classifier Alignment for Long-Tail Class Incremental Learning

Jayateja Kalla, Soma Biswas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 32-41

This paper introduces a two-stage framework designed to enhance long-tail class incremental learning, enabling the model to progressively learn new classes, while mitigating catastrophic forgetting in the context of long-tailed data distributions. Addressing the challenge posed by the under-representation of tail classes in long-tail class incremental learning, our approach achieves classifier alignment by leveraging global variance as an informative measure and class prototypes in the second stage. This process effectively captures class properties and eliminates the need for data balancing or additional layer tuning. Alongside traditional class incremental learning losses in the first stage, the proposed approach incorporates mixup classes to learn robust feature representations, ensurin g smoother boundaries. The proposed framework can seamlessly integrate as a module with any class incremental learning method to effectively handle long-tail class incremental learning scenarios. Extensive experimentation on the CIFAR-100 and ImageNet-Subset datasets validates the approach's efficacy, showcasing its superiority over state-of-the-art techniques across various long-tail CIL settings.

****

MagneticPillars: Efficient Point Cloud Registration Through Hierarchized Birds-Eye-View Cell Correspondence Refinement

Kai Fischer, Martin Simon, Stefan Milz, Patrick Mäder; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7386-7395

Recent point cloud registration approaches often deal with a consecutive determination of coarse and fine feature correspondences for hierarchical pose refinement. Due to the unordered nature of point clouds, a common way to generate a subsampled representation for the coarse matching step is by applying 3D-sensitive convolution approaches. However, expensive grouping mechanisms such as nearest neighbour search have to be used to determine the associated fine features, generating individual associations for each point cloud and leading to an increased overall runtime. Furthermore current methods often tend to predict deficient point correspondences and rely on additional filtering by expensive registration backends like RANSAC impeding their application in time critical systems. To overcome these challenges, we present MagneticPillars utilizing a Birds-Eye-View (BEV) grid representation, entailing fixed affiliations between coarse and fine feature cells. We show that by extracting correspondences in this manner, a small amount of key points is already sufficient to achieve an accurate pose estimation without external optimization methods like RANSAC. We evaluate our approach on two autonomous driving datasets for the task of point cloud registration by applying SVD as the backend, where we outperform recent state-of-the-art methods, reducing the rotation and translation error by 12% and 40%, respectively, and to top it all off, cutting runtime in half.

****

Fast Diffusion EM: A Diffusion Model for Blind Inverse Problems With Application to Deconvolution

Charles Laroche, Andrés Almansa, Eva Coupeté; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5271-5281

Using diffusion models to solve inverse problems is a growing field of research. Current methods assume the degradation to be known and provide impressive results in terms of restoration quality and diversity. In this work, we leverage the efficiency of those models to jointly estimate the restored image and unknown parameters of the degradation model such as blur kernel. In particular, we designed an algorithm based on the well-known Expectation-Minimization (EM) estimation method and diffusion models. Our method alternates between approximating the expected log-likelihood of the inverse problem using samples drawn from a diffusion model and a maximization step to estimate unknown model parameters. For the maximization step, we also introduce a novel blur kernel regularization based on a Plug & Play denoiser. Diffusion models are long to run, thus we provide a fast version of our algorithm. Extensive experiments on blind image deblurring demonstrate the effectiveness of our method when compared to other state-of-the-art approaches.

****

Active Transfer Learning for Efficient Video-Specific Human Pose Estimation

Hiromu Taketsugu, Norimichi Ukita; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1880-1890

Human Pose (HP) estimation is actively researched because of its wide range of applications. However, even estimators pre-trained on large datasets may not perform satisfactorily due to a domain gap between the training and test data. To address this issue, we present our approach combining Active Learning (AL) and Transfer Learning (TL) to adapt HP estimators to individual video domains efficiently. For efficient learning, our approach quantifies (i) the estimation uncertain ty based on the temporal changes in the estimated heatmaps and (ii) the unnaturalness in the estimated full-body HPs. These quantified criteria are then effectively combined with the state-of-the-art representativeness criterion to select uncertain and diverse samples for efficient HP estimator learning. Furthermore, we reconsider the existing Active Transfer Learning (ATL) method to introduce novel ideas related to the retraining methods and Stopping Criteria (SC). Experimental results demonstrate that our method enhances learning efficiency and outperforms comparative methods. Our code is publicly available at: https://github.com/ImIntheMiddle/VATL4Pose-WACV2024

****

Training-Free Layout Control With Cross-Attention Guidance

Minghao Chen, Iro Laina, Andrea Vedaldi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5343-5353

Recent diffusion-based generators can produce high-quality images from textual prompts. However, they often disregard textual instructions that specify the spatial layout of the composition. We propose a simple approach that achieves robust layout control without the need for training or fine-tuning of the image generator. Our technique manipulates the cross-attention layers that the model uses to interface textual and visual information and steers the generation in the desired direction given, e.g., a user-specified layout. To determine how to best guide attention, we study the role of attention maps and explore two alternative strategies, forward and backward guidance. We thoroughly evaluate our approach on three benchmarks and provide several qualitative examples and a comparative analysis of the two strategies that demonstrate the superiority of backward guidance compared to forward guidance, as well as prior work. We further demonstrate the versatility of layout guidance by extending it to applications such as editing the layout and context of real images.

****

Learning Transferable Representations for Image Anomaly Localization Using Dense Pretraining

Haitian He, Sarah Erfani, Mingming Gong, Qiuhong Ke; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1113-1122

Image anomaly localization (IAL) is widely applied in fault detection and industrial inspection domains to discover anomalous patterns in images at the pixel level. The unique challenge of this task is the lack of comprehensive anomaly samples for model training. The state-of-the-art methods train end-to-end models that leverage outlier exposure to simulate pseudo anomalies, but they show poor transferability to new datasets due to the inherent bias to the synthesized outliers during training. Recently, two-stage instance-level self-supervised learning (SSL) has shown potential in learning generic representations for IAL. However, we hypothesize that dense-level SSL is more compatible as IAL requires pixel-level prediction. In this paper, we bridge these gaps by proposing a two-stage, dense pre-training model tailored for the IAL task. More specifically, our model utilizes dual positive-pair selection criteria and dual feature scales to learn more effective representations. Through extensive experiments, we show that our learned representations achieve significantly better anomaly localization performance among two-stage models, while requiring almost half the convergence time. Moreover, our learned representations have better transferability to unseen datasets. Code is available at https://github. com/terrlo/DS2.

****

Embedding Task Structure for Action Detection

Michael Peven, Gregory D. Hager; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6604-6613

We present a straightforward, flexible method to enhance the accuracy and quality of action detection by expressing temporal and structural relationships of actions in the loss function of a deep network. We describe ways to represent otherwise implicit structure in video data and demonstrate how these structures reflect natural biases that improve network training. Our experiments show that our approach improves both accuracy and edit-distance of action recognition and detec tion models over a baseline. Our framework leads to improvements over prior work and obtains state-of-the-art results on multiple benchmarks.

****

RIMeshGNN: A Rotation-Invariant Graph Neural Network for Mesh Classification

Bahareh Shakibajahromi, Edward Kim, David E. Breen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3150-3160

Shape analysis tasks, including mesh classification, segmentation, and retrieval demonstrate symmetries in Euclidean space and should be invariant to geometric transformations such as rotation and translation. However, existing methods in mesh analysis often rely on extensive data augmentation and more complex analysis models to handle 3D rotations. Despite these efforts, rotation invariance is not guaranteed, which can significantly reduce accuracy when test samples undergo arbitrary rotations, because the analysis method struggles to generalize to the unknown orientations of the test samples. To address these challenges, our work presents a novel approach that employs graph neural networks (GNNs) to analyze mesh-structured data. Our proposed GNN layer, aggregation function, and local pooling layer are equivariant to the rotation, reflection and translation of 3D shapes, making them suitable building blocks for our proposed rotation-invariant network for the classification of mesh models. Therefore, our proposed approach does not need rotation augmentation, and we can maintain accuracy even when test samples undergo arbitrary rotations. Extensive experiments on various datasets demonstrate that our methods achieve state-of-the-art performance.

****

Stereo Matching in Time: 100+ FPS Video Stereo Matching for Extended Reality

Ziang Cheng, Jiayu Yang, Hongdong Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8719-8728

Real-time Stereo Matching is a cornerstone task for Extended Reality (XR) applications, such as 3D scene understanding, video pass-through, and mixed-reality games. Despite significant advancements, getting accurate depth information in real time on a low-power mobile device remains a challenge. One of the main difficulties is the lack of high-quality indoor video stereo data captured by head-mounted VR or AR glasses. To address this, we introduce a novel video stereo synthetic dataset that comprises photorealistic renderings of various indoor scenes and realistic camera motion captured by a moving VR/AR head-mounted display (HMD). Our newly proposed dataset enables one to develop a novel framework for continuous video-rate stereo matching. As another contribution, we also propose a new video-based stereo matching approach tailored for XR applications, which achieves real-time inference at an impressive 134fps on a standard desktop computer, or 30fps on a battery-powered HMD. Our key insight is that disparity and contextual information are highly correlated and redundant between consecutive stereo frames. By unrolling an iterative cost aggregation in time (i.e. in temporal dimension), we are able to distribute and reuse the aggregated features over time. This leads to a substantial reduction in computation without sacrificing accuracy. We conducted extensive evaluations and demonstrated that our method achieves superior performance compared to the current state-of-the-art, making it a strong contender for real-time stereo matching in VR/AR applications.

****

Learning the What and How of Annotation in Video Object Segmentation

Thanos Delatolas, Vicky Kalogeiton, Dim P. Papadopoulos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6951-6961

Video Object Segmentation (VOS) is crucial for several applications, from video editing to video data generation. Training a VOS model requires an abundance of manually labeled training videos. The de-facto traditional way of annotating objects requires humans to draw detailed segmentation masks on the target objects at each video frame. This annotation process, however, is tedious and time-consuming. To reduce this annotation cost, in this paper, we propose EVA-VOS, a human-in-the-loop annotation framework for video object segmentation. Unlike the traditional approach, we introduce an agent that predicts iteratively both which frame ("What") to annotate and which annotation type ("How") to use. Then, the annot ator annotates only the selected frame that is used to update a VOS module, leading to significant gains in annotation time. We conduct experiments on the MOSE and the DAVIS datasets and we show that: (a) EVA-VOS leads to masks with accuracy close to the human agreement 3.5x faster than the standard way of annotating videos; (b) our frame selection achieves state-of-the-art performance; (c) EVA-VOS yields significant performance gains in terms of annotation time compared to all other methods and baselines.

****

Reverse Knowledge Distillation: Training a Large Model Using a Small One for Retinal Image Matching on Limited Data

Sahar Almahfouz Nasser, Nihar Gupte, Amit Sethi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7778-7787

Retinal image matching (RIM) plays a crucial role in monitoring disease progression and treatment response as retina is the only tissue where blood vessels can be directly observed. However, datasets with matched keypoints between temporally separated pairs of images are not available in abundance to train transformer-based models. Firstly, we release keypoint annotations for retinal images from multiple datasets to aid further research on RIM. Secondly, we propose a novel approach based on reverse knowledge distillation to train large models with limited data while preventing overfitting. We propose architectural modifications to a CNN-based semi-supervised method called SuperRetina [22] that helps improve its results on a publicly available dataset. We train a computationally heavier model based on a vision transformer encoder, utilizing the lighter CNN-based model. This approach, which we call reverse knowledge distillation (RKD), further improves the matching results even though it contrasts with the conventional knowledge distillation where lighter models are trained based on heavier ones is the norm. Further, we show that our technique generalizes to other domains, such as facial landmark matching.

****

Edge Inference With Fully Differentiable Quantized Mixed Precision Neural Networks

Clemens JS Schaefer, Siddharth Joshi, Shan Li, Raul Blazquez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8460-8469

The large computing and memory cost of deep neural networks (DNNs) often precludes their use in resource-constrained devices. Quantizing the parameters and operations to lower bit-precision offers substantial memory and energy savings for neural network inference, facilitating the use of DNNs on edge computing platforms. Recent efforts at quantizing DNNs have employed a range of techniques encompassing progressive quantization, step-size adaptation, and gradient scaling. This paper proposes a new quantization approach for mixed precision convolutional neural networks (CNNs) targeting edge-computing. Our method establishes a new pareto frontier in model accuracy and memory footprint demonstrating a range of pre-trained quantized models, delivering best-in-class accuracy below 4.3 MB of weights and activations without modifying the model architecture. Our main contributions are: (i) a method for tensor-sliced learned precision with a hardware-aware cost function for heterogeneous differentiable quantization, (ii) targeted gradient modification for weights and activations to mitigate quantization errors, and (iii) a multi-phase learning schedule to address instability in learning arising from updates to the learned quantizer and model parameters. We demonstrate the effectiveness of our techniques on the ImageNet dataset across a range of models including EfficientNet-Lite0 (e.g., 4.14MB of weights and activations at 67.66% accuracy) and MobileNetV2 (e.g., 3.51MB weights and activations at 65.39% accuracy).

****

CAD - Contextual Multi-Modal Alignment for Dynamic AVQA

Asmar Nadeem, Adrian Hilton, Robert Dawes, Graham Thomas, Armin Mustafa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7251-7263

In the context of Audio Visual Question Answering (AVQA) tasks, the audio and vi sual modalities could be learnt on three levels: 1) Spatial, 2) Temporal, and 3) Semantic. Existing AVQA methods suffer from two major shortcomings; the audio-visual (AV) information passing through the network isn't aligned on Spatial and Temporal levels; and, inter-modal (audio and visual) Semantic information is often not balanced within a context; this results in poor performance. In this paper, we propose a novel end-to-end Contextual Multi-modal Alignment (CAD) network that addresses the challenges in AVQA methods by i) introducing a parameter-free stochastic Contextual block that ensures robust audio and visual alignment on the Spatial level; ii) proposing a pre-training technique for dynamic audio and visual alignment on Temporal level in a self-supervised setting, and iii) introducing a cross-attention mechanism to balance audio and visual information on Semantic level. The proposed novel CAD network improves the overall performance over the state-of-the-art methods on average by 9.4% on the MUSIC-AVQA dataset. We also demonstrate that our proposed contributions to AVQA can be added to the existing methods to improve their performance without additional complexity requirements.

****

Discriminator-Free Unsupervised Domain Adaptation for Multi-Label Image Classification

Inder Pal Singh, Enjie Ghorbel, Anis Kacem, Arunkumar Rathinam, Djamila Aouada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3936-3945

In this paper, a discriminator-free adversarial-based Unsupervised Domain Adaptation (UDA) for Multi-Label Image Classification (MLIC) referred to as DDA-MLIC is proposed. Recently, some attempts have been made for introducing adversarial-based UDA methods in the context of MLIC. However, these methods which rely on an additional discriminator subnet present one major shortcoming. The learning of domain-invariant features may harm their task-specific discriminative power, since the classification and discrimination tasks are decoupled. Herein, we propose to overcome this issue by introducing a novel adversarial critic that is directly deduced from the task-specific classifier. Specifically, a two-component Gaussian Mixture Model (GMM) is fitted on the source and target predictions in order to distinguish between two clusters. This allows extracting a Gaussian distribution for each component. The resulting Gaussian distributions are then used for formulating an adversarial loss based on a Frechet distance. The proposed method is evaluated on several multi-label image datasets covering three different types of domain shift. The obtained results demonstrate that DDA-MLIC outperforms existing state-of-the-art methods in terms of precision while requiring a lower number of parameters. The code is publicly available at github.com/cvi2snt/DDA-MLIC.

****

Continual Test-Time Domain Adaptation via Dynamic Sample Selection

Yanshuo Wang, Jie Hong, Ali Cheraghian, Shafin Rahman, David Ahmedt-Aristizabal, Lars Petersson, Mehrtash Harandi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1701-1710

The objective of Continual Test-time Domain Adaptation (CTDA) is to gradually adapt a pre-trained model to a sequence of target domains without accessing the source data. This paper proposes a Dynamic Sample Selection (DSS) method for CTDA. DSS consists of dynamic thresholding, positive learning, and negative learning processes. Traditionally, models learn from unlabeled unknown environment data and equally rely on all samples' pseudo-labels to update their parameters through self-training. However, noisy predictions exist in these pseudo-labels, so all samples are not equally trustworthy. Therefore, in our method, a dynamic thresholding module is first designed to select suspected low-quality from high-quality samples. The selected low-quality samples are more likely to be wrongly predicted. Therefore, we apply joint positive and negative learning on both high- and low-quality samples to reduce the risk of using wrong information. We conduct extensive experiments that demonstrate the effectiveness of our proposed method for CTDA in the image domain, outperforming the state-of-the-art results. Furthermore, our approach is also evaluated in the 3D point cloud domain, showcasing its

versatility and potential for broader applicability.
****

FuseCap: Leveraging Large Language Models for Enriched Fused Image Captions
Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, Ron Kimmel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5689-5700

The advent of vision-language pre-training techniques enhanced substantial progress in the development of models for image captioning. However, these models frequently produce generic captions and may omit semantically important image details. This limitation can be traced back to the image-text datasets; while their captions typically offer a general description of image content, they frequently omit salient details. Considering the magnitude of these datasets, manual reannotation is impractical, emphasizing the need for an automated approach. To address this challenge, we leverage existing captions and explore augmenting them with visual details using "frozen" vision experts including an object detector, an attribute recognizer, and an Optical Character Recognizer (OCR). Our proposed method, FuseCap, fuses the outputs of such vision experts with the original captions using a large language model (LLM), yielding comprehensive image descriptions. We automatically curate a training set of 12M image-enriched caption pairs. These pairs undergo extensive evaluation through both quantitative and qualitative analyses. Subsequently, this data is utilized to train a captioning generation BLIP-based model. This model outperforms current state-of-the-art approaches, producing more precise and detailed descriptions, demonstrating the effectiveness of the proposed data-centric approach. We release this large-scale dataset of enriched image-caption pairs for the community.
****

Learning To Adapt CLIP for Few-Shot Monocular Depth Estimation
Xueting Hu, Ce Zhang, Yi Zhang, Bowen Hai, Ke Yu, Zhihai He; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5594-5603

Pre-trained Visual-Language Models (VLMs), such as CLIP, have shown enhanced performance across a range of tasks that involve the integration of visual and linguistic elements. When CLIP is used for depth estimation tasks, the patches, divided from the input images, can be combined with a series of semantic descriptions of the depth information to obtain similarity results. The coarse estimation of depth is then achieved by weighting and summing the depth values, called depth bins, corresponding to the predefined semantic descriptions. The zero-shot approach circumvents the computational and time-intensive nature of traditional fully-supervised depth estimation methods. However, this method, utilizing fixed depth bins, may not effectively generalize as images from different scenes may exhibit distinct depth distributions. To address this challenge, we propose a few-shot-based method which learns to adapt the VLMs for monocular depth estimation to balance training costs and generalization capabilities. Specifically, it assigns different depth bins for different scenes, which can be selected by the model during inference. Additionally, we incorporate learnable prompts to preprocess the input text to convert the easily human-understood text into easily model-understood vectors and further enhance the performance. With only one image per scene for training, our extensive experiment results on the NYU V2 dataset demonstrate that our method outperforms the previous state-of-the-art method by up to 10.6% in terms of MARE.
****

Asymmetric Image Retrieval With Cross Model Compatible Ensembles
Alon Shoshan, Ori Linial, Nadav Bhonker, Elad Hirsch, Lior Zamir, Igor Kviatkovsky, Gérard Medioni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1-11

The asymmetrical retrieval setting is a well suited solution for resource constrained applications such as face recognition and image retrieval. In this setting, a large model is used for indexing the gallery while a lightweight model is used for querying. The key principle in such systems is ensuring that both models share the same embedding space. Most methods in this domain are based on knowled ge distillation. While useful, they suffer from several drawbacks: they are upper-bounded by the performance of the single best model found and cannot be extended to use an ensemble of models in a straightforward manner. In this paper we present an approach that does not rely on knowledge distillation, rather it utilizes embedding transformation models. This allows the use of N independently trained and diverse gallery models (e.g., trained on different datasets or having a different architecture) and a single query model. As a result, we improve the overall accuracy beyond that of any single model while maintaining a low computational budget for querying. Additionally, we propose a gallery image rejection method that utilizes the diversity between multiple transformed embeddings to estimate the uncertainty of gallery images.

****

Progressive Hypothesis Transformer for 3D Human Mesh Recovery

Huang-Ru Liao, Jen-Chun Lin, Chun-Yi Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6323-6332

Recent advancements in Transformer-based human mesh reconstruction (HMR) are commendable. However, these models often lift 2D images directly to 3D vertices without explicit intermediate guidance. In addition, the global attention mechanism tends to spread attention across larger body areas and even unrelated background regions during human mesh estimation, rather than focusing on critical local regions such as human body joints. This tendency leads to inaccurate and unrealistic results for complex activities. To address these challenges, we introduce the Progressive Hypotheses Transformer, which employs 2D and 3D pose predictions to progressively guide our model. Moreover, we propose a mechanism that generates multiple plausible hypotheses for both 2D and 3D poses to mitigate potential inaccuracies arising from intermediate pose estimations. Our model also incorporates inter-intra attention to capture correlations between joints and hypotheses. Experimental results demonstrate that our method surpasses existing imagebased approaches on Human3.6M [13] and 3DPW [36] with fewer parameters and relatively lower computational costs.

****

MPT: Mesh Pre-Training With Transformers for Human Pose and Mesh Reconstruction

Kevin Lin, Chung-Ching Lin, Lin Liang, Zicheng Liu, Lijuan Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3415-3425

Traditional methods of reconstructing 3D human pose and mesh from single images rely on paired image-mesh datasets, which can be difficult and expensive to obtain. Due to this limitation, model scalability is constrained as well as reconstruction performance. Towards addressing the challenge, we introduce Mesh Pre-Training (MPT), an effective pre-training strategy that leverages large amounts of MoCap data to effectively perform pre-training at scale. We introduce the use of MoCap-generated heatmaps as input representations to the mesh regression transformer and propose a Masked Heatmap Modeling approach for improving pre-training performance. This study demonstrates that pre-training using the proposed MPT allows our models to perform effective inference without requiring fine-tuning. We further show that fine-tuning the pre-trained MPT model considerably improves the accuracy of human mesh reconstruction from single images. Experimental results show that MPT outperforms previous state-of-the-art methods on Human3.6M and 3DPW datasets. As a further application, we benchmark and study MPT on the task of 3D hand reconstruction, showing that our generic pre-training scheme generalizes well to hand pose estimation and achieves promising reconstruction performance.

****

Training-Free Content Injection Using H-Space in Diffusion Models

Jaeseok Jeong, Mingi Kwon, Youngjung Uh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5151-5161

Diffusion models (DMs) synthesize high-quality images in various domains. However, controlling their generative process is still hazy because the intermediate variables in the process are not rigorously studied. Recently, the bottleneck feature of the U-Net, namely h-space, is found to convey the semantics of the resul ting image. It enables StyleCLIP-like latent editing within DMs. In this paper, we explore further usage of h-space beyond attribute editing, and introduce a method to inject the content of one image into another image by combining their features in the generative processes. Briefly, given the original generative process of the other image, 1) we gradually blend the bottleneck feature of the content with proper normalization, and 2) we calibrate the skip connections to match the injected content. Unlike custom-diffusion approaches, our method does not require time-consuming optimization or fine-tuning. Instead, our method manipulates intermediate features within a feed-forward generative process. Furthermore, our method does not require supervision from external networks.

****

Hard Sample-Aware Consistency for Low-Resolution Facial Expression Recognition

Bokyeung Lee, Kyungdeuk Ko, Jonghwan Hong, Hanseok Ko; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 199-208

Facial expression recognition (FER) plays a pivotal role in computer vision applications, encompassing video understanding and human-computer interaction. Despite notable advancements in FER, performance still falters when handling low-resolution facial images encountered in real-world scenarios and datasets. While consistency constraint techniques have garnered attention for generating robust convolutional neural network models that accommodate input variations through augmentation, their efficacy is diminished in the realm of low-resolution FER. This decline in performance can be attributed to augmented samples that networks struggle to extract expressive features. In this paper, we identify hard samples that cause an overfitting problem when considering various degrees of resolution and propose novel hard sample-aware consistency (HSAC) loss functions, which include combined attention consistency and label distribution learning. The combined attention consistency aligns an attention map from multi-scale low-resolution images with an appropriate target attention map by combining activation maps from high-resolution and flipped low-resolution images. We measure the classification difficulty for low-resolution face images and adaptively apply label distribution learning by combining the original target and predictions of high-resolution input. Our HSAC empowers the network to achieve generalization by effectively managing hard samples. Extensive experiments on various FER datasets demonstrate the superiority of our proposed method over existing approaches for multi-scale low-resolution images. Furthermore, we achieved a new state-of-the-art performance of 90.97% on the original RAF-DB dataset.

****

ZEETAD: Adapting Pretrained Vision-Language Model for Zero-Shot End-to-End Temporal Action Detection

Thinh Phan, Khoa Vo, Duy Le, Gianfranco Doretto, Donald Adjeroh, Ngan Le; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7046-7055

Temporal action detection (TAD) involves the localization and classification of action instances within untrimmed videos. While standard TAD follows fully supervised learning with closed-set setting on large training data, recent zero-shot TAD methods showcase the promising openset setting by leveraging large-scale contrastive visuallanguage (ViL) pretrained models. However, existing zeroshot TAD methods have limitations on how to properly construct the strong relationship between two Interdependent tasks of localization and classification and adapt ViL model to video understanding. In this work, we present ZEETAD, featuring two modules: dual-localization and zeroshot proposal classification. The former is a Transformerbased module that detects action events while selectively collecting crucial semantic embeddings for later Recognition. The latter one, CLIP-based module, generates semantic embeddings from text and frame inputs for each temporal unit. Additionally, we enhance discriminative capability on unseen classes by minimally updating the frozen CLIP encoder with lightweight adapters. Extensive experiments on THUMOS14 and ActivityNet-1.3 datasets demonstrate our approach's superior performance in zero-shot TAD and effective knowledge transfer from ViL models to unseen action categories. Code is available at https: //github.com/UARK-A

ICV/ZEETAD.

****

Army of Thieves: Enhancing Black-Box Model Extraction via Ensemble Based Sample Selection

Akshit Jindal, Vikram Goyal, Saket Anand, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3823-3832

Machine Learning (ML) models become vulnerable to Model Stealing Attacks (MSA) when they are deployed as a service. In such attacks, the deployed model is queried repeatedly to build a labelled dataset. This dataset allows the attacker to train a thief model that mimics the original model. To maximize query efficiency, the attacker has to select the most informative subset of data points from the pool of available data. Existing attack strategies utilize approaches like Active Learning and Semi-Supervised learning to minimize costs. However, in the black-box setting, these approaches may select sub-optimal samples as they train only one thief model. Depending on the thief model's capacity and the data it was pretrained on, the model might even select noisy samples that harm the learning process. In this work, we explore the usage of an ensemble of deep learning models as our thief model. We call our attack Army of Thieves(AOT) as we train multiple models with varying complexities to leverage the crowd's wisdom. Based on the ensemble's collective decision, uncertain samples are selected for querying, while the most confident samples are directly included in the training data. Our approach is the first one to utilize an ensemble of thief models to perform model extraction. We outperform the base approaches of existing state-of-the-art methods by at least 3% and achieve a 21% higher adversarial sample transferability than previous work for models trained on the CIFAR-10 dataset.

****

GC-MVSNet: Multi-View, Multi-Scale, Geometrically-Consistent Multi-View Stereo

Vibhas K. Vats, Sripad Joshi, David J. Crandall, Md. Alimoor Reza, Soon-heung Jung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3242-3252

Traditional multi-view stereo (MVS) methods rely heavily on photometric and geometric consistency constraints, but newer machine learning-based MVS methods check geometric consistency across multiple source views only as a post-processing step. In this paper, we present a novel approach that explicitly encourages geometric consistency of reference view depth maps across multiple source views at different scales during learning (see Fig. 1). We find that adding this geometric consistency loss significantly accelerates learning by explicitly penalizing geometrically inconsistent pixels, reducing the training iteration requirements to nearly half that of other MVS methods. Our extensive experiments show that our approach achieves a new state-of-the-art on the DTU and BlendedMVS datasets, and competitive results on the Tanks and Temples benchmark. To the best of our knowledge, GC-MVSNet is the first attempt to enforce multi-view, multi-scale geometric consistency during learning.

****

Active Batch Sampling for Multi-Label Classification With Binary User Feedback

Debanjan Goswami, Shayok Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2534-2543

Multi-label classification is a generalization of multi-class classification, where a single data sample can have multiple labels. While deep neural networks have depicted commendable performance for multi-label learning, they require a large amount of manually annotated training data to attain good generalization capability. However, annotating a multi-label data sample requires a human oracle to consider the presence/absence of every single class individually, which is extremely laborious. Active learning algorithms automatically identify the salient and exemplar instances from large amounts of unlabeled data and are effective in reducing human annotation effort in inducing a machine learning model. In this paper, we propose a novel active learning framework for multi-label learning, which queries a batch of (image-label) pairs and for each pair, poses the question whether the queried label is present in the corresponding image; the human annot ators merely need to provide a binary feedback (yes / no) in response to each query, which involves much less manual work. We pose the image and label selection as a constrained optimization problem and derive a linear programming relaxation to select a batch of (image-label) pairs, which are maximally informative to the underlying deep neural network. Our extensive empirical studies on three challenging datasets corroborate the potential of our method for real-world multi-label classification applications.

****

Efficient MAE Towards Large-Scale Vision Transformers

Qiu Han, Gongjie Zhang, Jiaxing Huang, Peng Gao, Zhang Wei, Shijian Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 606-615

Masked Autoencoder (MAE) has demonstrated superb pre-training efficiency for vision Transformer, thanks to its partial input paradigm and high mask ratio (0.75). However, MAE often suffers from severe performance drop under higher mask ratios, which hinders its potential toward larger-scale vision Transformers. In this work, we identify that the performance drop is largely attributed to the over-dominance of difficult reconstruction targets, as higher mask ratios lead to more sparse visible patches and fewer visual clues for reconstruction. To mitigate this issue, we design Efficient MAE that introduces a novel Difficulty-Flatten Loss and a decoder masking strategy, enabling a higher mask ratio for more efficient pre-training. The Difficulty-Flatten Loss provides balanced supervision on reconstruction targets of different difficulties, mitigating the performance drop under higher mask ratios effectively. Additionally, the decoder masking strategy discards the most difficult reconstruction targets, which further alleviates the optimization difficulty and accelerates the pre-training clearly. Our proposed Efficient MAE introduces 27% and 30% pre-training runtime accelerations for the ViT-Large and ViT-Huge models, provides valuable insights into MAE's optimization, and paves the way for larger-scale vision Transformer pre-training. Code and pre-trained models will be released.

****

M33D: Learning 3D Priors Using Multi-Modal Masked Autoencoders for 2D Image and Video Understanding

Muhammad Abdullah Jamal, Omid Mohareri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2544-2554

We present a new pre-training strategy called M^ 3 3D (Multi-Modal Masked 3D) built based on Multi-modal masked autoencoders that can leverage 3D priors and learned cross-modal representations in RGB-D data. We integrate two major self-supervised learning frameworks; Masked Image Modeling (MIM) and contrastive learning; aiming to effectively embed masked 3D priors and modality complementary features to enhance the correspondence between modalities. In contrast to recent approaches which are either focusing on specific downstream tasks or require multi-view correspondence, we show that our pre-training strategy is ubiquitous, enabling improved representation learning that can transfer into improved performance on various downstream tasks such as video action recognition, video action detection, 2D semantic segmentation and depth estimation. Experiments show that M^ 3 3D outperforms the existing state-of-the-art approaches on ScanNet, NYUv2, UCF-101 and OR-AR, particularly with an improvement of +1.3% mIoU against Mask3D on ScanNet semantic segmentation. We further evaluate our method on low-data regime and demonstrate its superior data efficiency compared to current state-of-the-art approaches.

****

Graph(Graph): A Nested Graph-Based Framework for Early Accident Anticipation

Nupur Thakur, PrasanthSai Gouripeddi, Baoxin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7533-7541

Anticipating traffic accidents early using dashcam videos is an important task for ensuring road safety and building reliable intelligent autonomous vehicles. However, factors like high traffic on the roads, different types of accidents, limited angles of vision, etc. make this task very challenging. Using the early frames, a lot of existing methods predict a large number of false positives which poses a huge risk for all vehicles on the road. In this paper, we propose a novel end-to-end learning, nested graph-based framework named Graph(Graph) for early accident anticipation. It uses interactions between the objects in the same as well as the neighboring frames along with the global features to make precise predictions as early as possible. This way it is able to embed the local as well as global temporal information into the extracted features. Graph(Graph) outperforms state-of-the-art methods on different datasets by a large margin demonstrating its effectiveness. With empirical evidence, we highlight the importance of each component in Graph(Graph) and show their effect on the final performance. Our code is available at https://github.com/thakurnupur/Graph-Graph.

****

Iterative Multi-Granular Image Editing Using Diffusion Models

K. J. Joseph, Prateksha Udhayanan, Tripti Shukla, Aishwarya Agarwal, Srikrishna Karanam, Koustava Goswami, Balaji Vasan Srinivasan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8107-8116

Recent advances in text-guided image synthesis has dramatically changed how creative professionals generate artistic and aesthetically pleasing visual assets. To fully support such creative endeavors, the process should possess the ability to: 1) iteratively edit the generations and 2) control the spatial reach of desired changes (global, local or anything in between). We formalize this pragmatic problem setting as Iterative Multi-granular Editing. While there has been substantial progress with diffusion-based models for image synthesis and editing, they are all one shot (i.e., no iterative editing capabilities) and do not naturally yield multi-granular control (i.e., covering the full spectrum of local-to-global edits). To overcome these drawbacks, we propose EMILIE: Iterative Multi-granular Image Editor. EMILIE introduces a novel latent iteration strategy, which re-purposes a pre-trained diffusion model to facilitate iterative editing. This is complemented by a gradient control operation for multi-granular control. We introduce a new benchmark dataset to evaluate our newly proposed setting. We conduct exhaustive quantitatively and qualitatively evaluation against recent state-of-the-art approaches adapted to our task, to being out the mettle of EMILIE. We hope our work would attract attention to this newly identified, pragmatic problem setting.

****

Efficient Feature Distillation for Zero-Shot Annotation Object Detection

Zhuoming Liu, Xuefeng Hu, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 893-902

We propose a new setting for detecting unseen objects called Zero-shot Annotation object Detection (ZAD). It expands the zero-shot object detection setting by allowing the novel objects to exist in the training images and restricts the additional information the detector uses to novel category names. Recently, to detect unseen objects, largescale vision-language models (e.g., CLIP) are leveraged by different methods. The distillation-based methods have good overall performance but suffer from a long training schedule caused by two factors. First, existing work creates distillation regions biased to the base categories, which limits the distillation of novel category information. Second, directly using the raw feature from CLIP for distillation neglects the domain gap between the training data of CLIP and the detection datasets, which makes it difficult to learn the mapping from the image region to the vision-language feature space. To solve these problems, we propose Efficient feature distillation for Zero-shot Annotation object Detection (EZAD). Firstly, EZAD adapts the CLIP's feature space to the target detection domain by re-normalizing CLIP; Secondly, EZAD uses CLIP to generate distillation proposals with potential novel category names to avoid the distillation being overly biased toward the base categories. Finally, EZAD takes advantage of semantic meaning for regression to further improve the model performance. As a result, EZAD outperforms the previous distillation-based methods in COCO by 4% with a much shorter training schedule and achieves a 3% improvement on the LVIS dataset. Our code is available at https://github.com/dragonlzm/EZAD

****

SpectralCLIP: Preventing Artifacts in Text-Guided Style Transfer From a Spectral

Perspective

Zipeng Xu, Songlong Xing, Enver Sangineto, Nicu Sebe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5121-5130

Owing to the power of vision-language foundation models, e.g., CLIP, the area of image synthesis has seen recent important advances. Particularly, for style transfer, CLIP enables transferring more general and abstract styles without collecting the style images in advance, as the style can be efficiently described with natural language, and the result is optimized by minimizing the CLIP similarity between the text description and the stylized image. However, directly using CLIP to guide style transfer leads to undesirable artifacts (mainly written words and unrelated visual entities) spread over the image. In this paper, we propose SpectralCLIP, which is based on a spectral representation of the CLIP embedding sequence, where most of the common artifacts occupy specific frequencies. By masking the band including these frequencies, we can condition the generation process to adhere to the target style properties (e.g., color, texture, paint stroke, etc.) while excluding the generation of larger-scale structures corresponding to the artifacts. Experimental results show that SpectralCLIP prevents the generation of artifacts effectively in quantitative and qualitative terms, without impairing the stylisation quality. We also apply SpectralCLIP to text-conditioned image generation and show that it prevents written words in the generated images. Our code is available at https://github.com/zipengxuc/SpectralCLIP.

****

Harnessing the Power of Multi-Lingual Datasets for Pre-Training: Towards Enhancing Text Spotting Performance

Alloy Das, Sanket Biswas, Ayan Banerjee, Josep Lladós, Umapada Pal, Saumik Bhattacharya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 718-728

The adaptation capability to a wide range of domains is crucial for scene text spotting models when deployed to real-world conditions. However, existing state-of-the-art approaches usually incorporate scene text detection and recognition simply by pretraining on natural scene image datasets, which do not directly exploit the feature interaction between multiple domains. In this work, we investigate the problem of domain-adapted scene text spotting, i.e., training a model on multi-domain source data such that it can directly adapt to target domains rather than being specialized for a specific domain or scenario. Further, we investigate a transformer baseline called Swin-TESTR to focus on solving scene-text spotting for both regular (ICDAR2015) and arbitrary-shaped scene text (CTW1500, TotalText) along with an exhaustive evaluation. The results clearly demonstrate the potential of intermediate representations on text spotting benchmarks across multiple domains (e.g. language, synth to real, and documents) both in terms of accuracy and model efficiency.

****

Rethink Cross-Modal Fusion in Weakly-Supervised Audio-Visual Video Parsing

Yating Xu, Conghui Hu, Gim Hee Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5615-5624

Existing works on weakly-supervised audio-visual video parsing adopt hybrid attention network (HAN) as the multi-modal embedding to capture the cross-modal context. It embeds the audio and visual modalities with a shared network, where the cross-attention is performed at the input. However, such an early fusion method highly entangles the two non-fully correlated modalities and leads to sub-optimal performance in detecting single-modality events. To deal with this problem, we propose the messenger-guided mid-fusion transformer to reduce the uncorrelated cross-modal context in the fusion. The messengers condense the full cross-modal context into a compact representation to only preserve useful cross-modal information. Furthermore, due to the fact that microphones capture audio events from all directions, while cameras only record visual events within a restricted field of view, there is a more frequent occurrence of unaligned cross-modal context from audio streams for visual event predictions. We thus propose cross-audio prediction consistency to suppress the impact of irrelevant audio information on vis ual event prediction. Experiments consistently illustrate the superior performance of our framework compared to existing state-of-the-art methods.

****

Refine and Redistribute: Multi-Domain Fusion and Dynamic Label Assignment for Unbiased Scene Graph Generation

Yujie Zang, Yaochen Li, Yuan Gao, Yimou Guo, Wenneng Tang, Yanxue Li, Meklit Atlaw; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1318-1327

Scene Graph Generation (SGG) plays an important role in enhancing visual image comprehension. However, existing approaches often struggle to represent implicit relationship features, resulting in a limited ability to distinguish predicates. Meanwhile, they are vulnerable to skewed instance distributions, which impairs effective training for fine-grained predicates. To address these problems, we propose a novel feature refinement and data redistribution framework (RAR). Specifically, a multi-domain fusion (MDF) module is designed to acquire comprehensive predicate representations, integrating global knowledge from the contextual domain and local details in the spatial-frequency domains. Then, we introduce a dynamic label assignment (DLA) strategy to tackle the long-tailed problem. Different predicate categories are adaptively grouped, accommodating varying training conditions. Guided by this strategy, we leverage a hierarchical auto-encoder to generate siamese samples, expanding the label cardinality. Furthermore, we explore the updated sample space to derive reliable samples and assign tailored labels, ultimately achieving the data rebalancing. Experiments on VG and GQA demonstrate that our model contributes to correcting prediction bias and achieves a significant improvement of approximately 10% in mean recall compared to baseline models.

****

Semantic Transfer From Head to Tail: Enlarging Tail Margin for Long-Tailed Visual Recognition

Shan Zhang, Yao Ni, Jinhao Du, Yanxia Liu, Piotr Koniusz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1350-1360

Deep neural networks excel in visual recognition tasks,but their success hinges on access to balanced datasets. Yet, real-world datasets often exhibit a long-tailed distribution, compromising network efficiency and hampering generalization on unseen data. To enhance the model's generalization in long-tailed scenarios, we present a novel feature augmentation approach termed SeMAntic tRansfer from head to Tail (SMART), which enriches the feature patterns for tail samples by transferring semantic covariance from the head classes to the tail classes along semantically correlating dimensions. This strategy boosts the model's generalization ability by implicitly and adaptively weighting the logits, thereby widening the classification margin of tail classes. Inspired by the success of this weighting, we further incorporate a semantic-aware weighting strategy for the loss tied to tail samples. This amplifies the effect of enlarging the margin for tail classes. We are the first to provide theoretical analysis that demonstrates a large semantic diversity in tail samples can increase class margins during the training stage, leading to improved generalization. Empirical observations support our theory. Notably, with no need for extra data or learnable parameters, SMART achieves state-of-the-art results on five long-tailed benchmark datasets: CIFAR-10/100-LT, Places-LT, ImageNet-LT, and iNaturalist 2018.

****

PoseDiff: Pose-Conditioned Multimodal Diffusion Model for Unbounded Scene Synthesis From Sparse Inputs

Seoyoung Lee, Joonseok Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5007-5017

Novel view synthesis has been heavily driven by NeRF-based models, but these models often hold limitations with the requirement of dense coverage of input views and expensive computations. NeRF models designed for scenarios with a few sparse input views face difficulty in being generalizable to complex or unbounded scenes, where multiple scene content can be at any distance from a multi-directiona

l camera, and thus generate unnatural and low quality images with blurry or floating artifacts. To accommodate the lack of dense information in sparse view scenarios and the computational burden of NeRF-based models in novel view synthesis, our approach adopts diffusion models. In this paper, we present PoseDiff, which combines the fast and plausible generation ability of diffusion models and 3D-aware view consistency of pose parameters from NeRF-based models. Specifically, PoseDiff is a multimodal pose-conditioned diffusion model applicable for novel view synthesis of unbounded scenes as well as bounded or forward-facing scenes with sparse views. PoseDiff renders plausible novel views for given pose parameters while maintaining high-frequency geometric details in significantly less time than conventional NeRF-based methods.
****
Leveraging Task-Specific Pre-Training To Reason Across Images and Videos
Arka Sadhu, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5794-5804
We explore the Reasoning Across Images and Video (RAIV) task, which requires models to reason on a pair of visual inputs comprising various combinations of images and/or videos. Previous work in this area has been limited to image pairs focusing primarily on the existence and/or cardinality of objects. To address this, we leverage existing datasets with rich annotations to generate semantically meaningful queries about actions, objects, and their relationships. We introduce new datasets that encompass visually similar inputs, reasoning over images, across images and videos, or across videos. Recognizing the distinct nature of RAIV compared to existing pre-training objectives which work on single image-text pairs, we explore task-specific pre-training, wherein a pre-trained model is trained on an objective similar to downstream tasks without utilizing fine-tuning datasets. Experiments with several state-of-the-art pre-trained image-language models reveal that task-specific pre-training significantly enhances performance on downstream datasets, even in the absence of additional pre-training data. We provide further ablative studies to guide future work.
****
Recognition of Unseen Bird Species by Learning From Field Guides
Andrés C. Rodríguez, Stefano D'Aronco, Rodrigo Caye Daudt, Jan D. Wegner, Konrad Schindler; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1742-1751
We exploit field guides to learn bird species recognition, in particular zero-shot recognition of unseen species. Illustrations contained in field guides deliberately focus on discriminative properties of each species, and can serve as side information to transfer knowledge from seen to unseen bird species. We study two approaches: (1) a contrastive encoding of illustrations, which can be fed into standard zero-shot learning schemes; and (2) a novel method that leverages the fact that illustrations are also images and as such structurally more similar to photographs than other kinds of side information. Our results show that illustrations from field guides, which are readily available for a wide range of species, are indeed a competitive source of side information for zero-shot learning. On a subset of the iNaturalist2021 dataset with 749 seen and 739 unseen species, we obtain a classification accuracy of unseen bird species of 12% @top-1 and 38% @top-10, which shows the potential of field guides for challenging real-world scenarios with many species. Our code is available at https://github.com/ac-rodriguez/zsl_billow.
****
LidarCLIP or: How I Learned To Talk to Point Clouds
Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, Lennart Svensson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7438-7447
Research connecting text and images has recently seen several breakthroughs, with models like CLIP, DALL*E 2, and Stable Diffusion. However, the connection between text and other visual modalities, such as lidar data, has received less attention, prohibited by the lack of text-lidar datasets. In this work, we propose LidarCLIP, a mapping from automotive point clouds to a pre-existing CLIP embeddin

g space. Using image-lidar pairs, we supervise a point cloud encoder with the image CLIP embeddings, effectively relating text and lidar data with the image domain as an intermediary. We show the effectiveness of LidarCLIP by demonstrating that lidar-based retrieval is generally on par with image-based retrieval, but with complementary strengths and weaknesses. By combining image and lidar features, we improve upon both single-modality methods and enable a targeted search for challenging detection scenarios under adverse sensor conditions. We also explore zero-shot classification and show that LidarCLIP outperforms existing attempts to use CLIP for point clouds by a large margin. Finally, we leverage our compatibility with CLIP to explore a range of applications, such as point cloud captioning and lidar-to-image generation, without any additional training. Code and pre-trained models at https://github.com/atonderski/lidarclip.

****

Enhancing Multimodal Compositional Reasoning of Visual Language Models With Generative Negative Mining

Ugur Sahin, Hang Li, Qadeer Khan, Daniel Cremers, Volker Tresp; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5563-5573

Contemporary large-scale visual language models (VLMs) exhibit strong representation capacities, making them ubiquitous for enhancing the image and text understanding tasks. They are often trained in a contrastive manner on a large and diverse corpus of images and corresponding text captions scraped from the internet. Despite this, VLMs often struggle with compositional reasoning tasks which require a fine-grained understanding of the complex interactions of objects and their attributes. This failure can be attributed to two main factors: 1) Contrastive approaches have traditionally focused on mining negative examples from existing datasets. However, the mined negative examples might not be difficult for the model to discriminate from the positive. An alternative to mining would be negative sample generation 2) But existing generative approaches primarily focus on generating hard negative texts associated with a given image. Mining in the other direction, i.e., generating negative image samples associated with a given text has been ignored. To overcome both these limitations, we propose a framework that not only mines in both directions but also generates challenging negative samples in both modalities, i.e., images and texts. Leveraging these generative hard negative samples, we significantly enhance VLMs' performance in tasks involving multimodal compositional reasoning. Our code and dataset are released at https://ugorsahin.github.io/enhancing-multimodal-compositional-reasoning-of-vlm.html.

****

LaughTalk: Expressive 3D Talking Head Generation With Laughter

Kim Sung-Bin, Lee Hyun, Da Hye Hong, Suekyeong Nam, Janghoon Ju, Tae-Hyun Oh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6404-6413

Laughter is a unique expression, essential to affirmative social interactions of humans. Although current 3D talking head generation methods produce convincing verbal articulations, they often fail to capture the vitality and subtleties of laughter and smiles despite their importance in social context. In this paper, we introduce a novel task to generate 3D talking heads capable of both articulate speech and authentic laughter. Our newly curated dataset comprises 2D laughing videos paired with pseudo-annotated and human-validated 3D FLAME parameters and vertices. Given our proposed dataset, we present a strong baseline with a two-stage training scheme: the model first learns to talk and then acquires the ability to express laughter. Extensive experiments demonstrate that our method performs favorably compared to existing approaches in both talking head generation and expressing laughter signals. We further explore potential applications on top of our proposed method for rigging realistic avatars.

****

Effects of Markers in Training Datasets on the Accuracy of 6D Pose Estimation

Janis Rosskamp, Rene Weller, Gabriel Zachmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4457-4466

Collecting training data for pose estimation methods on images is a time-consumi ng task and usually involves some kind of manual labeling of the 6D pose of objects. This time could be reduced considerably by using marker-based tracking that would allow for automatic labeling of training images. However, images containing markers may reduce the accuracy of pose estimation due to a bias introduced by the markers. In this paper, we analyze the influence of markers in training images on pose estimation accuracy. We investigate the accuracy of estimated poses for three different cases: i) training on images with markers, ii) removing markers by inpainting, and iii) augmenting the dataset with randomly generated markers to reduce spatial learning of marker features. Our results demonstrate that utilizing marker-based techniques is an effective strategy for collecting large amounts of ground truth data for pose prediction. Moreover, our findings suggest that the usage of inpainting techniques do not reduce prediction accuracy. Additionally, we investigate the effect of inaccuracies of labeling in training data on prediction accuracy. We show that the precise ground truth data obtained through marker tracking proves to be superior compared to markerless datasets if labeling errors of 6D ground truth exist. Our data generation tools are available online: https://github.com/JHRosskamp/6DPoseDataGenTools

****

Alleviating Foreground Sparsity for Semi-Supervised Monocular 3D Object Detection

Weijia Zhang, Dongnan Liu, Chao Ma, Weidong Cai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7542-7552

Monocular 3D object detection (M3OD) is a significant yet inherently challenging task in autonomous driving due to absence of explicit depth cues in a single RGB image. In this paper, we strive to boost currently underperforming monocular 3D object detectors by leveraging an abundance of unlabelled data via semi-supervised learning. Our proposed ODM3D framework entails cross-modal knowledge distillation at various levels to inject LiDAR-domain knowledge into a monocular detector during training. By identifying object sparsity as the main culprit behind existing methods' suboptimal training, we exploit the precise localisation information embedded in LiDAR points to enable more foreground-attentive and efficient distillation via the proposed BEV occupancy guidance mask, leading to notably improved knowledge transfer and M3OD performance. Besides, motivated by insights into why existing cross-modal GT-sampling techniques fail on our task at hand, we further design a novel cross-modal object-wise data augmentation strategy for effective RGB-LiDAR joint learning. Our method ranks 1st in both KITTI validation and test benchmarks, significantly surpassing all existing monocular methods, supervised or semi-supervised, on both BEV and 3D detection metrics.

****

MFT: Long-Term Tracking of Every Pixel

Michal Neoral, Jonáš Šerých, Ji■í Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6837-6847

We propose MFT -- Multi-Flow dense Tracker -- a novel method for dense, pixel-level, long-term tracking. The approach exploits optical flows estimated not only between consecutive frames, but also for pairs of frames at logarithmically spaced intervals. It selects the most reliable sequence of flows on the basis of estimates of its geometric accuracy and the probability of occlusion, both provided by a pre-trained CNN. We show that MFT achieves competitive performance on the TAP-Vid benchmark, outperforming baselines by a significant margin, and tracking densely orders of magnitude faster than the state-of-the-art point-tracking methods. The method is insensitive to medium-length occlusions and it is robustified by estimating flow with respect to the reference frame, which reduces drift.

****

Out-of-Distribution Detection With Logical Reasoning

Konstantin Kirchheim, Tim Gonschorek, Frank Ortmeier; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2122-2131

Machine Learning models often only generalize reliably to samples from the training distribution. Consequentially, detecting when input data is out-of-distribution (OOD) is crucial, especially in safety-critical applications. Current OOD de tection methods, however, tend to be domain agnostic and often fail to incorporate valuable prior knowledge about the structure of the training distribution. To address this limitation, we introduce a novel, hybrid OOD detection algorithm that combines a deep learning-based perception system with a first-order logic-based knowledge representation. A logical reasoning system uses this knowledge base at run-time to infer whether inputs are consistent with prior knowledge about the training distribution. In contrast to purely neural systems, the structured knowledge representation allows humans to inspect and modify the rules that govern the OOD detectors' behavior. This not only enhances performance but also fosters a level of explainability that is particularly beneficial in safety-critical contexts. We demonstrate the effectiveness of our method through experiments on several datasets and discuss advantages and limitations. Our code is available online.

****

WalkFormer: Point Cloud Completion via Guided Walks

Mohang Zhang, Yushi Li, Rong Chen, Yushan Pan, Jia Wang, Yunzhe Wang, Rong Xiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3293-3302

Point clouds are often sparse and incomplete in real-world scenarios. The prevailing methods for point cloud completion typically rely on encoding the partial points and then decoding complete points from a global feature vector, which might lose the existing patterns and elaborate structures. To address these issues, we propose WalkFormer, a novel approach to predict complete point clouds through a partial deformation process. Concretely, our method samples locally dominant points based on feature similarity and moves the points to form the missing part. Since these points maintain representative information of the surrounding structures, they are appropriately selected as the starting points for multiple guided walks. Furthermore, we design a Route Transformer module to exploit and aggregate the walk information with topological relations. These guided walks facilitate the learning of long-range dependencies for predicting shape deformation. Qualitative and quantitative evaluations demonstrate that our proposed approach achieves superior performance compared to state-of-the-art methods in the 3D point cloud completion task.

****

Driving Through the Concept Gridlock: Unraveling Explainability Bottlenecks in Automated Driving

Jessica Echterhoff, An Yan, Kyungtae Han, Amr Abdelraouf, Rohit Gupta, Julian McAuley; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7346-7355

Concept bottleneck models have been successfully used for explainable machine learning by encoding information within the model with a set of human-defined concepts. In the context of human-assisted or autonomous driving, explainability models can help user acceptance and understanding of decisions made by the autonomous vehicle, which can be used to rationalize and explain driver or vehicle behavior. We propose a new approach using concept bottlenecks as visual features for control command predictions and explanations of user and vehicle behavior. We learn a human understandable concept layer that we use to explain sequential driving scenes while learning vehicle control commands. This approach can then be used to determine whether a change in a preferred gap or steering commands from a human (or autonomous vehicle) is led by an external stimulus or change in preferences. We achieve competitive performance to latent visual features while gaining interpretability within our model setup.

****

Single-Image Deblurring, Trajectory and Shape Recovery of Fast Moving Objects With Denoising Diffusion Probabilistic Models

Radim Spetlik, Denys Rozumnyi, Ji■í Matas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6857-6866

Blurry appearance of fast moving objects in video frames was successfully used to reconstruct the object appearance and motion in both 2D and 3D domains. The proposed method addresses the novel, severely ill-posed, task of single-image fast moving object deblurring, shape, and trajectory recovery -- previous approaches require at least three consecutive video frames. Given a single image, the method outputs the object 2D appearance and position in a series of sub-frames as if captured by a high-speed camera (i.e. temporal super-resolution). The proposed SI-DDPM-FMO method is trained end-to-end on a synthetic dataset with various moving objects, yet it generalizes well to real-world data from several publicly available datasets. SI-DDPM-FMO performs similarly to or better than recent multi-frame methods and a carefully designed baseline method.

****

IDD-AW: A Benchmark for Safe and Robust Segmentation of Drive Scenes in Unstructured Traffic and Adverse Weather

Furqan Ahmed Shaik, Abhishek Reddy, Nikhil Reddy Billa, Kunal Chaudhary, Sunny Manchanda, Girish Varma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4614-4623

Large-scale deployment of fully autonomous vehicles requires a very high degree of robustness to unstructured traffic, weather conditions, and should prevent unsafe mispredictions. While there are several datasets and benchmarks focusing on segmentation for drive scenes, they are not specifically focused on safety and robustness issues. We introduce the IDD-AW dataset, which provides 5000 pairs of high-quality images with pixel-level annotations, captured under rain, fog, low light, and snow in unstructured driving conditions. As compared to other adverse weather datasets, we provide i.) more annotated images, ii.) paired Near-Infrared (NIR) image for each frame, iii.) larger label set with a 4-level label hierarchy to capture unstructured traffic conditions. We benchmark state-of-the-art models for semantic segmentation in IDD-AW. We also propose a new metric called "Safe mean Intersection over Union (Safe mIoU)" for hierarchical datasets which penalizes dangerous mispredictions that are not captured in the traditional definition of mean Intersection over Union (mIoU). The results show that IDD-AW is one of the most challenging datasets to date for these tasks. The dataset and code will be available here: https://iddaw.github.io.

****

Semantic Generative Augmentations for Few-Shot Counting

Perla Doubinsky, Nicolas Audebert, Michel Crucianu, Hervé Le Borgne; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5443-5452

With the availability of powerful text-to-image diffusion models, recent works have explored the use of synthetic data to improve image classification performances. These works show that it can effectively augment or even replace real data. In this work, we investigate how synthetic data can benefit few-shot class-agnostic counting. This requires to generate images that correspond to a given input number of objects. However, text-to-image models struggle to grasp the notion of count. We propose to rely on a double conditioning of Stable Diffusion with both a prompt and a density map in order to augment a training dataset for few-shot counting. Due to the small dataset size, the fine-tuned model tends to generate images close to the training images. We propose to enhance the diversity of synthesized images by exchanging captions between images thus creating unseen configurations of object types and spatial layout. Our experiments show that our diversified generation strategy significantly improves the counting accuracy of two recent and performing few-shot counting models on FSC147 and CARPK.

****

Text-to-Image Models for Counterfactual Explanations: A Black-Box Approach

Guillaume Jeanneret, Loïc Simon, Frédéric Jurie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4757-4767

This paper addresses the challenge of generating Counterfactual Explanations (CEs), involving the identification and modification of the fewest necessary features to alter a classifier's prediction for a given image. Our proposed method, Text-to-Image Models for Counterfactual Explanations (TIME), is a black-box counterfactual technique based on distillation. Unlike previous methods, this approach requires solely the image and its prediction, omitting the need for the classifier's structure, parameters, or gradients. Before generating the counterfactuals

, TIME introduces two distinct biases into Stable Diffusion in the form of textual embeddings: the context bias, associated with the image's structure, and the class bias, linked to class-specific features learned by the target classifier. After learning these biases, we find the optimal latent code applying the classifier's predicted class token and regenerate the image using the target embedding as conditioning, producing the counterfactual explanation. Extensive empirical studies validate that TIME can generate explanations of comparable effectiveness even when operating within a black-box setting.

****

Physical-Space Multi-Body Mesh Detection Achieved by Local Alignment and Global Dense Learning

Haoye Dong, Tiange Xiang, Sravan Chittupalli, Jun Liu, Dong Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1267-1276

From monocular RGB images captured in the wild, detecting multi-body 3D meshes in physical sizes and locations is notoriously difficult due to the diverse visual ambiguity and lack of explicit depth measurement. Modern DNN approaches made numerous advances based on either two-stage Region-of-Interests(RoI)-Align or single-stage fixed Field-of-View (FoV) detector frameworks for two main subtasks: local pelvis-centered mesh regression and global body-to-camera translation regression. However, sub-meter-level physical-space monocular mesh detection is still out of reach by existing solutions. In this paper, we recognize two common drawbacks: (1) The local meshes are usually estimated without explicitly aligning body features under image-space scaling, occlusion, and truncation; (2) The global translations are estimated based on a weak-perspective assumption, which tricks the network into prioritizing image-space (front-view) mesh alignment and leads to inaccurate mesh depth. We introduce Physical-space Multi-body Mesh Detection (PMMD), in which (1) Locally, we preserve the body aspect ratio, align the body-to-RoI layout, and densely refine the person-wise RoI features for robustness; (2) Globally, we learn dense-depth-guided features to amend the body-wise local feature for physical depth estimation. With the cleaned local features and explicit local-global associations, PMMD achieves the best centimeter-level local mesh metrics and the first sub-meter-level global mesh metrics from monocular images in 3DPW and AGORA datasets.

****

Guided Cluster Aggregation: A Hierarchical Approach to Generalized Category Discovery

Jona Otholt, Christoph Meinel, Haojin Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2618-2627

Despite advances in image recognition, recognizing novel categories in unlabeled data remains challenging for machine learning methods, even though humans can perform this task with ease. A recently developed setting to tackle this problem is Generalized Category Discovery (GCD), in which the task is to, given a labeled dataset, classify an unlabeled dataset, where the unlabeled dataset contains both known classes and novel classes that do not appear in the labeled data. Existing GCD methods mostly focus on learning strong image representations, on which they then apply a clustering algorithm such as k-means. Despite obtaining good performance, they do not fully exploit the potential of the learned features due to the simple nature of the clustering mechanism. To address this issue, we make use of the fact that local neighborhoods in self-supervised feature spaces are highly homogeneous. We leverage this observation to develop Guided Cluster Aggregation (GCA), a hierarchical approach that first groups the data into small clusters of high purity, then aggregates them into larger clusters. Experiments show that GCA outperforms semi-supervised k-means in most cases, especially in fine-grained classification tasks. Code available at https://github.com/J- L- O/guided-cluster-aggregation.

****

Masked Event Modeling: Self-Supervised Pretraining for Event Cameras

Simon Klenk, David Bonello, Lukas Koestler, Nikita Araslanov, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (

WACV), 2024, pp. 2378-2388

Event cameras asynchronously capture brightness changes with low latency, high temporal resolution, and high dynamic range. However, annotation of event data is a costly and laborious process, which limits the use of deep learning methods for classification and other semantic tasks with the event modality. To reduce the dependency on labeled event data, we introduce Masked Event Modeling (MEM), a self-supervised framework for events. Our method pretrains a neural network on unlabeled events, which can originate from any event camera recording. Subsequently, the pretrained model is finetuned on a downstream task, leading to a consistent improvement of the task accuracy. For example, our method reaches state-of-the-art classification accuracy across three datasets, N-ImageNet, N-Cars, and N-Caltech101, increasing the top-1 accuracy of previous work by significant margins. When tested on real-world event data, MEM is even superior to supervised RGB-based pretraining. The models pretrained with MEM are also label-efficient and generalize well to the dense task of semantic image segmentation.

****

Real-Time Polyp Detection in Colonoscopy Using Lightweight Transformer

Youngbeom Yoo, Jae Young Lee, Dong-Jae Lee, Jiwoon Jeon, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7809-7819

Colorectal cancer (CRC) represents a major global health challenge, and early detection of polyps is crucial in preventing its progression. Although colonoscopy is the gold standard for polyp detection, it has limitations, such as human error and missed detection rates. In response, computer-aided detection (CADe) systems have been developed to enhance the efficiency and accuracy of polyp detection. As deep learning gained prominence, the incorporation of Convolutional Neural Networks (CNNs) into CADe systems emerged as a breakthrough approach. However, CADe systems based on CNNs often demand significant computational resources, making them unsuitable for deployment in resource-constrained environments. To mitigate this, we propose a novel and lightweight polyp detection model that integrates a Transformer layer into the You Only Look Once (YOLO) architecture, focusing on optimizing the neck part responsible for feature fusion and rescaling. Our model demonstrates a substantial reduction in computational complexity and the number of parameters, without compromising detection performances. The lightweight model makes it accessible and feasibly deployable in medically underserved regions, serving a significant public interest by potentially expanding the reach of critical diagnostic tools for CRC prevention. By optimizing the architecture to reduce resource requirements while maintaining performance, our model becomes a practical solution to assist healthcare professionals in the real-time identification of polyps, even with resource-constraint devices.

****

Top-Down Beats Bottom-Up in 3D Instance Segmentation

Maksim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, Danila Rukhovich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3566-3574

Most 3D instance segmentation methods exploit a bottom-up strategy, typically including resource-exhaustive post-processing. For point grouping, bottom-up methods rely on prior assumptions about the objects in the form of hyperparameters, which are domain-specific and need to be carefully tuned. On the contrary, we address 3D instance segmentation with a TD3D: the pioneering cluster-free, fully-convolutional and entirely data-driven approach trained in an end-to-end manner. This is the first top-down method outperforming bottom-up approaches in 3D domain. With its straightforward pipeline, it performs outstandingly well on the standard benchmarks: ScanNet v2, its extension ScanNet200, and S3DIS. Besides, our method is much faster on inference than the current state-of-the-art grouping-based approaches: our flagship modification is 1.9x faster than the most accurate bottom-up method, while being more accurate, and our faster modification shows state-of-the-art accuracy running at 2.6x speed. Code is available at https://github.com/SamsungLabs/td3d.

****

Open-Set Object Detection by Aligning Known Class Representations

Hiran Sarkar, Vishal Chudasama, Naoyuki Onoe, Pankaj Wasnik, Vineeth N. Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 219-228

Open Set Object Detection (OSOD) has emerged as a contemporary research direction to address the detection of unknown objects. Recently, few works have achieved remarkable performance in the OSOD task by employing contrastive clustering to separate unknown classes. In contrast, we propose a new semantic clustering-based approach to facilitate a meaningful alignment of clusters in semantic space and introduce a class decorrelation module to enhance inter-cluster separation. Our approach further incorporates an object focus module to predict objectness scores, which enhances the detection of unknown objects. Further, we employ i) an evaluation technique that penalizes low-confidence outputs to mitigate the risk of misclassification of the unknown objects and ii) a new metric called HMP that combines known and unknown precision using harmonic mean. Our extensive experiments demonstrate that the proposed model achieves significant improvement on the MS-COCO & PASCAL VOC dataset for the OSOD task.

****

DR2: Disentangled Recurrent Representation Learning for Data-Efficient Speech Video Synthesis

Chenxu Zhang, Chao Wang, Yifan Zhao, Shuo Cheng, Linjie Luo, Xiaohu Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6204-6214

Although substantial progress has been made in audio-driven talking video synthesis, there still remain two major difficulties: existing works 1) need a long sequence of training dataset (>1h) to synthesize co-speech gestures, which causes a significant limitation on their applicability; 2) usually fail to generate long sequences, or can only generate long sequences without enough diversity. To solve these challenges, we propose a Disentangled Recurrent Representation Learning framework to synthesize long diversified gesture sequences with a short training video of around 2 minutes. In our framework, we first make a disentangled latent space assumption to encourage unpaired audio and pose combinations, which results in diverse "one-to-many" mappings in pose generation. Next, we apply a recurrent inference module to feed back the last generation as initial guidance to the next phase, enhancing the long-term video generation of full continuity and diversity. Comprehensive experimental results verify that our model can generate realistic synchronized full-body talking videos with training data efficiency.

****

EvDNeRF: Reconstructing Event Data With Dynamic Neural Radiance Fields

Anish Bhattacharya, Ratnesh Madaan, Fernando Cladera, Sai Vemprala, Rogerio Bonatti, Kostas Daniilidis, Ashish Kapoor, Vijay Kumar, Nikolai Matni, Jayesh K. Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5846-5855

We present EvDNeRF, a pipeline for generating event data and training an event-based dynamic NeRF, for the purpose of faithfully reconstructing eventstreams on scenes with rigid and non-rigid deformations that may be too fast to capture with a standard camera. Event cameras register asynchronous per-pixel brightness changes at MHz rates with high dynamic range, making them ideal for observing fast motion with almost no motion blur. Neural radiance fields (NeRFs) offer visual-quality geometric-based learnable rendering, but prior work with events has only considered reconstruction of static scenes. Our EvDNeRF can predict eventstreams of dynamic scenes from a static or moving viewpoint between any desired timestamps, thereby allowing it to be used as an event-based simulator for a given scene. We show that by training on varied batch sizes of events, we can improve test-time predictions of events at fine time resolutions, outperforming baselines that pair standard dynamic NeRFs with event generators. We release our simulated and real datasets, as well as code for multi-view event-based data generation and the training and evaluation of EvDNeRF models.

****

DISCO: Distributed Inference With Sparse Communications

Minghai Qin, Chao Sun, Jaco Hofmann, Dejan Vucinic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2432-2440

Deep neural networks (DNNs) have great potential to solve many real-world problems, but they usually require an extensive amount of computation and memory. It is of great difficulty to deploy a large DNN model to a single resource-limited device with small memory capacity. Distributed computing is a common approach to reduce single-node memory consumption and to accelerate the inference of DNN models. In this paper, we explore the "within-layer model parallelism", which distributes the inference of each layer into multiple nodes. In this way, the memory requirement can be distributed to many nodes, making it possible to use several edge devices to infer a large DNN model. Due to the dependency within each layer, data communications between nodes during this parallel inference can be a bottleneck when the communication bandwidth is limited. We propose a framework to train DNN models for Distributed Inference with Sparse Communications (DISCO). We convert the problem of selecting which subset of data to transmit between nodes into a model optimization problem, and derive models with both computation and communication reduction when each layer is inferred on multiple nodes. We show the benefit of the DISCO framework on a variety of CV tasks such as image classification, object detection, semantic segmentation, and image super resolution. The corresponding models include important DNN building blocks such as convolutions and transformers. For example, each layer of a ResNet-50 model can be distributively inferred across two nodes with 5x less data communications, almost half overall computations and less than half memory requirement for a single node, and achieve comparable accuracy to the original ResNet-50 model.

****

EmoStyle: One-Shot Facial Expression Editing Using Continuous Emotion Parameters

Bita Azari, Angelica Lim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6385-6394

Recent studies have achieved impressive results in face generation and editing of facial expressions. However, existing approaches either generate a discrete number of facial expressions or have limited control over the emotion of the output image. To overcome this limitation, we introduced EmoStyle, a method to edit facial expressions based on valence and arousal, two continuous emotional parameters that can specify a broad range of emotions. EmoStyle is designed to separate emotions from other facial characteristics and to edit the face to display a desired emotion. We employ the pre-trained generator from StyleGAN2, taking advantage of its rich latent space. We also proposed an adapted inversion method to be able to apply our system on out-of-StyleGAN2 domain images in a one-shot manner. The qualitative and quantitative evaluations show that our approach has the capability to synthesize a wide range of expressions to output high-resolution images.

****

FinderNet: A Data Augmentation Free Canonicalization Aided Loop Detection and Closure Technique for Point Clouds in 6-DOF Separation.

Sudarshan S. Harithas, Gurkirat Singh, Aneesh Chavan, Sarthak Sharma, Suraj Patni, Chetan Arora, Madhava Krishna; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8399-8408

We focus on the problem of LiDAR point cloud based loop detection (or Finding) and closure (LDC) for mobile robots. State-of-the-art (SOTA) methods directly generate learned embeddings from a given point cloud, require large data augmentation, and are not robust to wide viewpoint variations in 6 Degrees-of-Freedom (DOF). Moreover, the absence of strong priors in an unstructured point cloud leads to highly inaccurate LDC. In this original approach, we propose independent roll and pitch canonicalization of point clouds using a common dominant ground plane. We discretize the canonicalized point clouds along the axis perpendicular to the ground plane leads to images simi- lar to digital elevation maps (DEMs), which expose strong spatial priors in the scene. Our experiments show that LDC based on learnt embeddings from such DEMs is not only data efficient but also significantly more robust, and generalizable than the current SOTA. We report an (average precision for loop detection, mean absolute transla- tion/rotation error) im provement of (8.4, 16.7/5.43)% on the KITTI08 sequence, and (11.0, 34.0/25.4)% on GPR10 sequence, over the current SOTA. To further test the ro- bustness of our technique on point clouds in 6-DOF motion we create and opensource a custom dataset called Lidar- UrbanFly Dataset (LUF) which consists of point clouds ob- tained from a LiDAR mounted on a quadrotor. More details on our website https://gsc2001.github.io/FinderNet/

****

Distortion-Disentangled Contrastive Learning

Jinfeng Wang, Sifan Song, Jionglong Su, S. Kevin Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 75-85

Self-supervised learning is well known for its remarkable performance in representation learning and various downstream computer vision tasks. Recently, Positive-pair-Only Contrastive Learning (POCL) has achieved reliable performance without the need to construct positive-negative training sets. It reduces memory requirements by lessening the dependency on the batch size. The POCL method typically uses a single objective function to extract the distortion invariant representation (DIR), which describes the proximity of positive-pair representations affected by different distortions. This objective function implicitly enables the model to filter out or ignore the distortion variant representation (DVR) affected by different distortions. However, some recent studies have shown that proper use of DVR in contrastive can optimize the performance of models in some downstream domain-specific tasks. In addition, these POCL methods have been observed to be sensitive to augmentation strategies. To address these limitations, we propose a novel POCL framework named Distortion-Disentangled Contrastive Learning (DDCL) and a Distortion-Disentangled Loss (DDL). Our approach is the first to explicitly and adaptively disentangle and exploit the DVR inside the model and feature stream to improve the overall representation utilization efficiency, robustness, and representation ability. Experiments demonstrate our framework's superiority to Barlow Twins and Simsiam in terms of convergence, representation quality (Including transferability and generality), and robustness on several benchmark datasets.

****

Boosting Weakly Supervised Object Detection Using Fusion and Priors From Hallucinated Depth

Cagri Gungor, Adriana Kovashka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 739-748

Despite recent attention and exploration of depth for various tasks, it is still an unexplored modality for weakly-supervised object detection (WSOD). We propose an amplifier method for enhancing the performance of WSOD by integrating depth information. Our approach can be applied to any WSOD method based on multiple instance learning, without necessitating additional annotations or inducing large computational expenses. Our proposed method employs a monocular depth estimation technique to obtain hallucinated depth information, which is then incorporated into a Siamese WSOD network using contrastive loss and fusion. By analyzing the relationship between language context and depth, we calculate depth priors to identify the bounding box proposals that may contain an object of interest. These depth priors are then utilized to update the list of pseudo ground-truth boxes, or adjust the confidence of perbox predictions. Our proposed method is evaluated on six datasets (COCO, PASCAL VOC, Conceptual Captions, Clipart1k, Watercolor2k, and Comic2k) by implementing it on top of two state-of-the-art WSOD methods, and we demonstrate a substantial enhancement in performance.

****

MS-EVS: Multispectral Event-Based Vision for Deep Learning Based Face Detection

Saad Himmi, Vincent Parret, Ajad Chhatkuli, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 616-625

Event-based sensing is a relatively new imaging modality that enables low latency, low power, high temporal resolution and high dynamic range acquisition. These properties make it a highly desirable sensor for edge applications and in high dynamic range environments. As of today, most event-based sensors are monochroma tic (grayscale), capturing light from a wide spectral range over the visible, in a single channel. In this paper, we introduce multispectral events and study their advantages. In particular, we consider multiple bands in the visible and near-infrared range, and explore their potential compared to monochromatic events and conventional multispectral imaging for the face detection task. We further release the first large scale bimodal face detection datasets, with RGB videos and their simulated color events, N-MobiFace and N-YoutubeFaces, and a smaller dataset with multispectral videos and events, N-SpectralFace. We find that early fusion of multispectral events significantly improves the face detection performance, compared to the early fusion of conventional multispectral images. This result shows that polychromatic events carry relatively more useful information about the scene than conventional multispectral/color images do, with respect to their monochromatic equivalent. To the best of our knowledge, our proposed method is the first exploratory research on multispectral events, specifically including near infrared data.

****

Adaptive Latent Diffusion Model for 3D Medical Image to Image Translation: Multi-Modal Magnetic Resonance Imaging Study

Jonghun Kim, Hyunjin Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7604-7613

Multi-modal images play a crucial role in comprehensive evaluations in medical image analysis providing complementary information for identifying clinically important biomarkers. However, in clinical practice, acquiring multiple modalities can be challenging due to reasons such as scan cost, limited scan time, and safety considerations. In this paper, we propose a model based on the latent diffusion model (LDM) that leverages switchable blocks for image-to-image translation in 3D medical images without patch cropping. The 3D LDM combined with conditioning using the target modality allows generating high-quality target modality in 3D overcoming the shortcoming of the missing out-of-slice information in 2D generation methods. The switchable block, noted as multiple switchable spatially adaptive normalization (MS-SPADE), dynamically transforms source latents to the desired style of the target latents to help with the diffusion process. The MS-SPADE block allows us to have one single model to tackle many translation tasks of one source modality to various targets removing the need for many translation models for different scenarios. Our model exhibited successful image synthesis across different source-target modality scenarios and surpassed other models in quantitative evaluations tested on multi-modal brain magnetic resonance imaging datasets of four different modalities. Our model demonstrated successful image synthesis across various modalities even allowing for one-to-many modality translations. Furthermore, it outperformed other one-to-one translation models in quantitative evaluations.

****

Let's Observe Them Over Time: An Improved Pedestrian Attribute Recognition Approach

Kamalakar Vijay Thakare, Debi Prosad Dogra, Heeseung Choi, Haksub Kim, Ig-Jae Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 708-717

Despite poor image quality, occlusions, and small training datasets, recent pedestrian attribute recognition (PAR) methods have achieved considerable performance. However, leveraging only spatial information of different attributes limits their reliability and generalizability. This paper introduces a multi-perspective approach to reduce over-dependence on spatial clues of a single perspective and exploits other aspects available in multiple perspectives. In order to tackle image quality and occlusions, we exploit different spatial clues present across images and handpick the best attribute-specific features to classify. Precisely, we extract the class-activation energy of each attribute and correlate it with the corresponding energy present across other images using the proposed Self-Attentive Cross Relation Module. In the next stage, we fuse this correlation information with similar clues accumulated from the other images. Lastly, we train a classification neural network using combined correlation information with two diff erent losses. We have validated our method on four widely used PAR datasets, namely Market1501, PETA, PA-100k, and Duke. Our method achieves superior performance over most existing methods, demonstrating the effectiveness of a multi-perspective approach in PAR.

****

AnyStar: Domain Randomized Universal Star-Convex 3D Instance Segmentation

Neel Dey, Mazdak Abulnaga, Benjamin Billot, Esra Abaci Turk, Ellen Grant, Adrian V. Dalca, Polina Golland; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7593-7603

Star-convex shapes arise across bio-microscopy and radiology in the form of nuclei, nodules, metastases, and other units. Existing instance segmentation networks for such structures train on densely labeled instances for each dataset, which requires substantial and often impractical manual annotation effort. Further, significant reengineering or finetuning is needed when presented with new datasets and imaging modalities due to changes in contrast, shape, orientation, resolution, and density. We present AnyStar, a domain-randomized generative model that simulates synthetic training data of blob-like objects with randomized appearance, environments, and imaging physics to train general-purpose star-convex instance segmentation networks. As a result, networks trained using our generative model do not require annotated images from unseen datasets. A single network trained on our synthesized data accurately 3D segments C. elegans and P. dumerilii nuclei in fluorescence microscopy, mouse cortical nuclei in micro-CT, zebrafish brain nuclei in EM, and placental cotyledons in human fetal MRI, all without any retraining, finetuning, transfer learning, or domain adaptation. Code is available at https://github.com/neel-dey/AnyStar.

****

Solving the Plane-Sphere Ambiguity in Top-Down Structure-From-Motion

Lars Haalck, Benjamin Risse; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3485-3493

Drone-based land surveys and tracking applications with a moving camera require three-dimensional reconstructions from videos recorded using a downward facing camera and are usually generated by Structure-from-Motion (SfM) algorithms. Unfortunately, monocular SfM pipelines can fail in the presence of lens distortion due to a critical configuration resulting in a plane-sphere ambiguity which is characterized by severe curvatures of the reconstructions and erroneous relative camera pose estimations. We propose a 4-point minimal solver for the relative pose estimation for two views sharing the same radial distortion parameters (i.e. from the same camera) with a viewing direction perpendicular to the ground plane. To extract 3D reconstructions from continuous videos, the relative pose of pairwise frames is estimated by using the solver with RANSAC and the Sampson error where globally consistent distortion parameters are determined by taking the medial of all values. Moreover, we propose an additional regularizer for the final bundle adjustment to remove any remaining curvature of the reconstruction if necessary. We tested our methods on synthetic and real-world data and our results demonstrate a significant reduction of curvature and more accurate relative pose estimations. Our algorithm can be easily integrated into existing pipelines and is therefore a practical solution to resolve the plane-sphere ambiguity in a variety of top-down SfM applications.

****

PromptonomyViT: Multi-Task Prompt Learning Improves Video Transformers Using Synthetic Scene Data

Roei Herzig, Ofir Abramovich, Elad Ben Avraham, Assaf Arbelle, Leonid Karlinsky, Ariel Shamir, Trevor Darrell, Amir Globerson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6803-6815

Action recognition models have achieved impressive results by incorporating scene-level annotations, such as objects, their relations, 3D structure, and more. However, obtaining annotations of scene structure for videos requires a significant amount of effort to gather and annotate, making these methods expensive to train. In contrast, synthetic datasets generated by graphics engines provide powerful alternatives for generating scene-level annotations across multiple tasks. I n this work, we propose an approach to leverage synthetic scene data for improving video understanding. We present a multi-task prompt learning approach for video transformers, where a shared video transformer backbone is enhanced by a small set of specialized parameters for each task. Specifically, we add a set of "task prompts", each corresponding to a different task, and let each prompt predict task-related annotations. This design allows the model to capture information shared among synthetic scene tasks as well as information shared between synthetic scene tasks and a real video downstream task throughout the entire network. We refer to this approach as "Promptonomy", since the prompts model task-related structure. We propose the PromptonomyViT model (PViT), a video transformer that incorporates various types of scene-level information from synthetic data using the "Promptonomy" approach. PViT shows strong performance improvements on multiple video understanding tasks and datasets. Project page: https://ofir1080.github.io/PromptonomyViT/

****

Improving the Leaking of Augmentations in Data-Efficient GANs via Adaptive Negative Data Augmentation

Zhaoyu Zhang, Yang Hua, Guanxiong Sun, Hui Wang, Seán McLoone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5412-5421

Data augmentation (DA) has shown its effectiveness in training Data-Efficient GANs (DE-GANs). However, applying DA in DE-GANs results in transforming the distributions of generated data and real data to augmented distributions of generated data and real data. This augmentation process could produce some out-of-distribution samples, known as the leaking of augmentations problem, which is highly undesirable in DE-GANs training. Although some methods propose "leaking-free" DAs for DE-GANs, we theoretically and practically argue that the leaking of augmentations problem still exists in these methods. To alleviate the leaking of augmentations in DE-GANs, in this paper, we propose a simple yet effective method called adaptive negative data augmentation (ANDA) for DE-GANs, with a negligible computational cost increase. Specifically, ANDA adaptively augments the augmented distribution of generated data using the augmented distribution of negative real data, where the negative real data is produced by applying negative data augmentation (NDA) on the real data. In this case, potential leaking samples can be presented as "fake" instances to the discriminator adaptively, which avoids the generator (G) learning such samples, thus resulting in better performance. Extensive experiments on several datasets with different DE-GANs demonstrate that ANDA can effectively alleviate the leaking of augmentations problem during training and achieve better performance. Codes are available at https://github.com/zzhang05/ANDA

****

Enhancing Diverse Intra-Identity Representation for Visible-Infrared Person Re-Identification

Sejun Kim, Soonyong Gwon, Kisung Seo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2513-2522

Visible-Infrared person Re-Identification (VI-ReID) is a challenging task due to modality discrepancy. To reduce modality-gap, existing methods primarily focus on sample diversity, such as data augmentation or generating intermediate modality between Visible and Infrared. However, these methods do not consider the increase in intra-instance variance caused by sample diversity, and they focus on dominant features, which results in a remaining modality gap for hard samples. This limitation hinders performance improvement. We propose Intra-identity Representation Diversification (IRD) based metric learning to handle the intra-instance variance. Specifically IRD method enlarge the Intra-modality Intra-identity Representation Space (IIRS) for each modality within the same identity to learn diverse feature representation abilities. This enables the formation of a shared space capable of representing common features across hetero-modality, thereby reducing the modality gap more effectively. In addition, we introduce a HueGray (HG) data augmentation method, which increases sample diversity simply and effectively. Finally, we propose the Diversity Enhancement Network (DEN) for robustly hand ling intra-instance variance. The proposed method demonstrates superior performance compared to the state-of-the-art methods on the SYSU-MM01 and RegDB datasets. Notably, on the challenging SYSU-MM01 dataset, our approach achieves remarkable results with a Rank-1 accuracy of 76.36% and a mean Average Precision (mAP) of 71.30%.

****

Synergizing Contrastive Learning and Optimal Transport for 3D Point Cloud Domain Adaptation

Siddharth Katageri, Arkadipta De, Chaitanya Devaguptapu, VSSV Prasad, Charu Sharma, Manohar Kaul; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2942-2951

Recently, the fundamental problem of unsupervised domain adaptation (UDA) on 3D point clouds has been motivated by a wide variety of applications in robotics, virtual reality, and scene understanding, to name a few. The point cloud data acquisition procedures manifest themselves as significant domain discrepancies and geometric variations among both similar and dissimilar classes. The standard domain adaptation methods developed for images do not directly translate to point cloud data because of their complex geometric nature. To address this challenge, we leverage the idea of multimodality and alignment between distributions. We propose a new UDA architecture for point cloud classification that benefits from multimodal contrastive learning to get better class separation in both domains individually. Further, the use of optimal transport (OT) aims at learning source and target data distributions jointly to reduce the cross-domain shift and provide a better alignment. We conduct a comprehensive empirical study on PointDA-10 and GraspNetPC-10 and show that our method achieves state-of the-art performance on GraspNetPC-10 (with approx. 4-12% margin) and best average performance on PointDA-10. Our ablation studies and decision boundary analysis also validate the significance of our contrastive learning module and OT alignment.

****

Video Instance Matting

Jiachen Li, Roberto Henschel, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6668-6677

Conventional video matting outputs one alpha matte for all instances appearing in a video frame so that individual instances are not distinguished. While video instance segmentation provides time-consistent instance masks, results are unsatisfactory for matting applications, especially due to applied binarization. To remedy this deficiency, we propose Video Instance Matting (VIM), that is, estimating the alpha mattes of each instance at each frame of a video sequence. To tackle this challenging problem, we present MSG-VIM, a Mask Sequence Guided Video Instance Matting neural network, as a novel baseline model for VIM. MSG-VIM leverages a mixture of mask augmentations to make predictions robust to inaccurate and inconsistent mask guidance. It incorporates temporal mask and temporal feature guidance to improve the temporal consistency of alpha matte predictions. Furthermore, we build a new benchmark for VIM, called VIM50, which comprises 50 video clips with multiple human instances as foreground objects. To evaluate performances on the VIM task, we introduce a suitable metric called Video Instance-aware Matting Quality (VIMQ). Our proposed model MSG-VIM sets a strong baseline on the VIM50 benchmark and outperforms existing methods by a large margin.

****

DPPMask: Masked Image Modeling With Determinantal Point Processes

Junde Xu, Zikai Lin, Donghao Zhou, Yaodong Yang, Xiangyun Liao, Qiong Wang, Bian Wu, Guangyong Chen, Pheng-Ann Heng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2266-2276

Masked Image Modeling (MIM) has achieved impressive representative performance with the aim of reconstructing randomly masked images. Despite the empirical success, most previous works have neglected the important fact that it is unreasonable to force the model to reconstruct something beyond recovery, such as those masked objects. In this work, we show that uniformly random masking widely used in previous works unavoidably loses some key objects and changes original semantic information, resulting in a misalignment problem and hurting the representative learning eventually. To address this issue, we augment MIM with a new masking strategy namely the DPPMask by substituting the random process with Determinantal Point Process (DPPs) to reduce the semantic change of the image after masking. Our method is simple yet effective and requires no extra learnable parameters when implemented within various frameworks. In particular, we evaluate our method on two representative MIM frameworks, MAE and iBOT. We show that DPPMask surpassed random sampling under both lower and higher masking ratios, indicating that DPPMask makes the reconstruction task more reasonable. We further test our method on the background challenge and multi-class classification tasks, showing that our method is more robust at various tasks.

****

ShadowSense: Unsupervised Domain Adaptation and Feature Fusion for Shadow-Agnostic Tree Crown Detection From RGB-Thermal Drone Imagery

Rudraksh Kapil, Seyed Mojtaba Marvasti-Zadeh, Nadir Erbilgin, Nilanjan Ray; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8266-8276

Accurate detection of individual tree crowns from remote sensing data poses a significant challenge due to the dense nature of forest canopy and the presence of diverse environmental variations, e.g., overlapping canopies, occlusions, and varying lighting conditions. Additionally, the lack of data for training robust models adds another limitation in effectively studying complex forest conditions. This paper presents a novel method for detecting shadowed tree crowns and provides a challenging dataset comprising roughly 50k paired RGB-thermal images to facilitate future research for illumination-invariant detection. The proposed method (ShadowSense) is entirely self-supervised, leveraging domain adversarial training without source domain annotations for feature extraction and foreground feature alignment for feature pyramid networks to adapt domain-invariant representations by focusing on visible foreground regions, respectively. It then fuses complementary information of both modalities to effectively improve upon the predictions of an RGB-trained detector and boost the overall accuracy. Extensive experiments demonstrate the superiority of the proposed method over both the baseline RGB-trained detector and state-of-the-art techniques that rely on unsupervised domain adaptation or early image fusion. Our code and data are available: https://github.com/rudrakshkapil/ShadowSense

****

Pruning From Scratch via Shared Pruning Module and Nuclear Norm-Based Regularization

Donghyeon Lee, Eunho Lee, Youngbae Hwang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1393-1402

Most pruning methods focus on determining redundant channels from the pre-trained model. However, they overlook the cost of training large networks and the significance of selecting channels for effective reconfiguration. In this paper, we present a "pruning from scratch" framework that considers reconfiguration and expression capacity. Our Shared Pruning Module (SPM) handles a channel alignment problem in residual blocks for lossless reconfiguration after pruning. Moreover, we introduce nuclear norm-based regularization to preserve the representability of large networks during the pruning process. By combining it with MACs-based regularization, we achieve an efficient and powerful pruned network while compressing towards target MACs. The experimental results demonstrate that our method prunes redundant channels effectively to enhance representation capacity of the network. Our approach compresses ResNet50 on ImageNet without requiring additional resources, achieving a top-1 accuracy of 75.25% with only 41% of the original model's MACs. Code is available at https://github.com/jsleeg98/NuSPM.

****

Semantic Labels-Aware Transformer Model for Searching Over a Large Collection of Lecture-Slides

K. V. Jobin, Anand Mishra, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6016-6025

Massive Open Online Courses (MOOCs) enable easy access to many educational mater ials, particularly lecture slides, on the web. Searching through them based on user queries becomes an essential problem due to the availability of such vast information. To address this, we present Lecture Slide Deck Search Engine -- a model that supports natural language queries and hand-drawn sketches and performs searches on a large collection of slide images on computer science topics. This search engine is trained using a novel semantic label-aware transformer model that extracts the semantic labels in the slide images and seamlessly encodes them with the visual cues from the slide images and textual cues from the natural language query. Further, to study the problem in a challenging setting, we introduce a novel dataset, namely the Lecture Slide Deck (LecSD) Dataset containing 54K slide images from the Data Structure, computer networks, and optimization courses and provide associated manual annotation for the query in the form of natural language or hand-drawn sketch. The proposed Lecture Slide Deck Search Engine outperforms the competitive baselines and achieves nearly 4% superior Recall@1 on an absolute scale compared to the state-of-the-art approach. We firmly believe that this work will open up promising directions for improving the accessibility and usability of educational resources, enabling students and educators to find and utilize lecture materials more effectively.

****

Multimodal Channel-Mixing: Channel and Spatial Masked AutoEncoder on Facial Action Unit Detection

Xiang Zhang, Huiyuan Yang, Taoyue Wang, Xiaotian Li, Lijun Yin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6077-6086

Recent studies have focused on utilizing multi-modal data to develop robust models for facial Action Unit (AU) detection. However, the heterogeneity of multi-modal data poses challenges in learning effective representations. One such challenge is extracting relevant features from multiple modalities using a single feature extractor. Moreover, previous studies have not fully explored the potential of multi-modal fusion strategies. In contrast to the extensive work on late fusion, there are limited investigations on early fusion for channel information exploration. This paper presents a novel multi-modal reconstruction network, named Multimodal Channel-Mixing (MCM), as a pre-trained model to learn robust representation for facilitating multi-modal fusion. The approach follows an early fusion setup, integrating a Channel-Mixing module, where two out of five channels are randomly dropped. The dropped channels then are reconstructed from the remaining channels using masked autoencoder. This module not only reduces channel redundancy, but also facilitates multi-modal learning and reconstruction capabilities, resulting in robust feature learning. The encoder is fine-tuned on a downstream task of automatic facial action unit detection. Pretraining experiments were conducted on BP4D+, followed by fine-tuning on BP4D and DISFA to assess the effectiveness and robustness of the proposed framework. The results demonstrate that our method meets and surpasses the performance of state-of-the-art baseline method.

****

ZIGNeRF: Zero-Shot 3D Scene Representation With Invertible Generative Neural Radiance Fields

Kanghyeok Ko, Minhyeok Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4986-4995

Generative Neural Radiance Fields (NeRFs) have demonstrated remarkable proficiency in synthesizing multi-view images by learning the distribution of a set of unposed images. Despite the aptitude of existing Generative NeRFs in generating 3D-consistent high-quality random samples within data distribution, the creation of a 3D representation of a singular input image remains a formidable challenge. In this manuscript, we introduce ZIGNeRF, an innovative model that executes zero-shot Generative Adversarial Network (GAN) inversion for the generation of multi-view images from a single out-of-distribution image. The model is underpinned by a novel inverter that maps out-of-domain images into the latent code of the generator manifold. Notably, ZIGNeRF is capable of disentangling the object from the background and executing 3D operations such as 360-degree rotation or depth a nd horizontal translation. The efficacy of our model is validated using multiple real-image datasets: Cats, AFHQ, CelebA, CelebA-HQ, and CompCars.

****

SLoSH: Set Locality Sensitive Hashing via Sliced-Wasserstein Embeddings

Yuzhe Lu, Xinran Liu, Andrea Soltoggio, Soheil Kolouri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2566-2576

Learning from set-structured data is an essential problem with many applications in machine learning and computer vision. This paper focuses on non-parametric and data-independent learning from set-structured data using approximate nearest neighbor (ANN) solutions, particularly locality-sensitive hashing. We consider the problem of set retrieval from an input set query. Such a retrieval problem requires: 1) an efficient mechanism to calculate the distances/dissimilarities between sets, and 2) an appropriate data structure for fast nearest-neighbor search. To that end, we propose to use Sliced-Wasserstein embedding as a computationally efficient set-2-vector operator that enables downstream ANN, with theoretical guarantees. The set elements are treated as samples from an unknown underlying distribution, and the Sliced-Wasserstein distance is used to compare sets. We demonstrate the effectiveness of our algorithm, denoted as Set Locality Sensitive Hashing (SLoSH), on various set retrieval datasets and compare our proposed embedding with standard set embedding approaches, including Generalized Mean (GeM) embedding/pooling, Featurewise Sort Pooling (FSPool), Covariance Pooling, and Wasserstein embedding and show consistent improvement in retrieval results.

****

StreamMapNet: Streaming Mapping Network for Vectorized Online HD Map Construction

Tianyuan Yuan, Yicheng Liu, Yue Wang, Yilun Wang, Hang Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7356-7365

High-Definition (HD) maps are essential for the safety of autonomous driving systems. While existing techniques employ camera images and onboard sensors to generate vectorized high-precision maps, they are constrained by their reliance on single-frame input. This approach limits their stability and performance in complex scenarios such as occlusions, largely due to the absence of temporal information. Moreover, their performance diminishes when applied to broader perception ranges. In this paper, we present StreamMapNet, a novel online mapping pipeline adept at long-sequence temporal modeling of videos. StreamMapNet employs multi-point attention and temporal information which empowers the construction of large-range local HD maps with high stability and further addresses the limitations of existing methods. Furthermore, we critically examine widely used online HD Map construction benchmark and datasets, Argoverse2 and nuScenes, revealing significant bias in the existing evaluation protocols. We propose to resplit the benchmarks according to geographical spans, promoting fair and precise evaluations. Experimental results validate that StreamMapNet significantly outperforms existing methods across all settings while maintaining an online inference speed of 14.2 FPS. Our code is available at https://github.com/yuantianyuan01/StreamMapNet.

****

Blurry Video Compression: A Trade-Off Between Visual Enhancement and Data Compression

Dawit Mureja Argaw, Junsik Kim, In So Kweon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4280-4290

Existing video compression (VC) methods primarily aim to reduce the spatial and temporal redundancies between consecutive frames in a video while preserving its quality. In this regard, previous works have achieved remarkable results on videos acquired under specific settings such as instant (known) exposure time and shutter speed which often result in sharp videos. However, when these methods are evaluated on videos captured under different temporal priors, which lead to degradations like motion blur and low frame rate, they fail to maintain the quality of the contents. In this work, we tackle the VC problem in a general scenario where a given video can be blurry due to predefined camera settings or dynamics i n the scene. By exploiting the natural trade-off between visual enhancement and data compression, we formulate VC as a min-max optimization problem and propose an effective framework and training strategy to tackle the problem. Extensive experimental results on several benchmark datasets confirm the effectiveness of our method compared to several state-of-the-art VC approaches.

****

Correlation-Aware Active Learning for Surgery Video Segmentation

Fei Wu, Pablo Márquez-Neila, Mingyi Zheng, Hedyeh Rafii-Tari, Raphael Sznitman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2010-2020

Semantic segmentation is a complex task that relies heavily on large amounts of annotated image data. How- ever, annotating such data can be time-consuming and resource-intensive, especially in the medical domain. Active Learning (AL) is a popular approach that can help to reduce this burden by iteratively selecting images for annotation to improve the model performance. In the case of video data, it is important to consider the model uncertainty and the temporal nature of the sequences when selecting images for annotation. This work proposes a novel AL strategy for surgery video segmentation, COWAL, COrrelation aWare Active Learning. Our approach involves projecting images into a latent space that has been fine-tuned using contrastive learning and then selecting a fixed number of representative images from local clusters of video frames. We demonstrate the effectiveness of this approach on two video datasets of surgical instruments and three real-world video datasets. The datasets and code will be made publicly available upon receiving necessary approvals.

****

EResFD: Rediscovery of the Effectiveness of Standard Convolution for Lightweight Face Detection

Joonhyun Jeong, Beomyoung Kim, Joonsang Yu, YoungJoon Yoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 988-998

This paper analyzes the design choices of face detection architecture that improve efficiency of computation cost and accuracy. Specifically, we re-examine the effectiveness of the standard convolutional block as a lightweight backbone architecture for face detection. Unlike the current tendency of lightweight architecture design, which heavily utilizes depthwise separable convolution layers, we show that heavily channel-pruned standard convolution layers can achieve better accuracy and inference speed when using a similar parameter size. This observation is supported by the analyses concerning the characteristics of the target data domain, faces. Based on our observation, we propose to employ ResNet with a highly reduced channel, which surprisingly allows high efficiency compared to other mobile-friendly networks (e.g., MobileNetV1, V2, V3). From the extensive experiments, we show that the proposed backbone can replace that of the state-of-the-art face detector with a faster inference speed. Also, we further propose a new feature aggregation method to maximize the detection performance. Our proposed detector EResFD obtained 80.4% mAP on WIDER FACE Hard subset which only takes 37.7 ms for VGA image inference on CPU. Code is available at https://github.com/clovaai/EResFD.

****

Neural Echos: Depthwise Convolutional Filters Replicate Biological Receptive Fields

Zahra Babaiee, Peyman M. Kiasari, Daniela Rus, Radu Grosu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8216-8225

In this study, we present evidence suggesting that depthwise convolutional kernels are effectively replicating the structural intricacies of the biological receptive fields observed in the mammalian retina. We provide analytics of trained kernels from various state-of-the-art models substantiating this evidence. Inspired by this intriguing discovery, we propose an initialization scheme that draws inspiration from the biological receptive fields. Experimental analysis of the ImageNet dataset with multiple CNN architectures featuring depthwise convolutions

reveals a marked enhancement in the accuracy of the learned model when initialized with biologically derived weights. This underlies the potential for biologically inspired computational models to further our understanding of vision processing systems and to improve the efficacy of convolutional networks.
****
Estimating Blood Alcohol Level Through Facial Features for Driver Impairment Assessment
Ensiyeh Keshtkaran, Brodie von Berg, Grant Regan, David Suter, Syed Zulqarnain Gilani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4539-4548
Drunk driving-related road accidents contribute significantly to the global burden of road injuries. Addressing alcohol-related harm, particularly during safety-critical activities like driving, requires real-time monitoring of an individual's blood alcohol concentration (BAC). We devise an in-vehicle machine learning system that harnesses standard commercial RGB cameras to predict critical levels of BAC. Our system can detect instances of alcohol intoxication impairment as subtle as 0.05 g/dL (WHO recommended legal limit for driving), with an accuracy of 75%, by leveraging the physiological manifestations of alcohol intoxication on a driver's face. This system holds great promise for improving road safety. In tandem, we have compiled a data set of 60 subjects engaged in simulated driving scenarios, spanning three levels of alcohol intoxication. These scenarios were captured and divided into video segments labeled "sober", "low", and "severe" Alcohol Intoxication Impairment (AII), constituting the basis for evaluating our system's performance. To the best of our knowledge, this study is the first to create a large-scale real-life dataset of alcohol intoxication and assess intoxication levels using an off-the-shelf RGB camera to detect drunk driving.
****
Auto-BPA: An Enhanced Ball-Pivoting Algorithm With Adaptive Radius Using Contextual Bandits
Houda Saffi, Naima Otberdout, Youssef Hmamouche, Amal El Fallah Seghrouchni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3729-3737
The Ball-Pivoting Algorithm (BPA) is a notable technique for 3D surface reconstruction from point clouds, heavily reliant on the ball radius. In practical application, determining the optimal radius for BPA often necessitates iterative experimentation to achieve better reconstruction quality. BPA entails geometric computations like iterative pivoting, inherently lacking differentiability. In this paper, we tackle the dual challenges of radius selection and non-differentiability in BPA. Inspired by contextual bandits, we propose an innovative approach that learns the optimal radius based on local geometric features within point clouds. We validate our method on the ModelNet10 and ShapeNet datasets, showcasing superior surface reconstruction compared to manual tuning and other classic methods both for low and high point cloud densities. Our code is available at https://github.com/houda-pixel/Auto-BPA.
****
MIDAS: Mixing Ambiguous Data With Soft Labels for Dynamic Facial Expression Recognition
Ryosuke Kawamura, Hideaki Hayashi, Noriko Takemura, Hajime Nagahara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6552-6562
Dynamic facial expression recognition (DFER) is an important task in the field of computer vision. To apply automatic DFER in practice, it is necessary to accurately recognize ambiguous facial expressions, which often appear in data in the wild. In this paper, we propose MIDAS, a data augmentation method for DFER, which augments ambiguous facial expression data with soft labels consisting of probabilities for multiple emotion classes. In MIDAS, the training data are augmented by convexly combining pairs of video frames and their corresponding emotion class labels, which can also be regarded as an extension of mixup to soft-labeled video data. This simple extension is remarkably effective in DFER with ambiguous facial expression data. To evaluate MIDAS, we conducted experiments on the DFEW

dataset. The results demonstrate that the model trained on the data augmented by MIDAS outperforms the existing state-of-the-art method trained on the original dataset.

****

MobileNVC: Real-Time 1080p Neural Video Compression on a Mobile Device

Ties van Rozendaal, Tushar Singhal, Hoang Le, Guillaume Sautiere, Amir Said, Krishna Buska, Anjuman Raha, Dimitris Kalatzis, Hitarth Mehta, Frank Mayer, Liang Zhang, Markus Nagel, Auke Wiggers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4323-4333

Neural video codecs have recently become competitive with standard codecs such as HEVC in the low-delay setting. However, most neural codecs are large floating-point networks that use pixel-dense warping operations for temporal modeling, making them too computationally expensive for deployment on mobile devices. Recent work has demonstrated that running a neural decoder in real time on mobile is feasible, but shows this only for 720p RGB video. This work presents the first neural video codec that decodes 1080p YUV420 video in real time on a mobile device. Our codec relies on two major contributions. First, we design an efficient codec that uses a block-based motion compensation algorithm available on the warping core of the mobile accelerator, and we show how to quantize this model to integer precision. Second, we implement a fast decoder pipeline that concurrently runs neural network components on the neural signal processor, parallel entropy coding on the mobile GPU, and warping on the warping core. Our codec outperforms the previous on-device codec by a large margin with up to 48 % BD-rate savings, while reducing the MAC count on the receiver side by 10x. We perform a careful ablation to demonstrate the effect of the introduced motion compensation scheme, and ablate the effect of model quantization.

****

Improving the Effectiveness of Deep Generative Data

Ruyu Wang, Sabrina Schmedding, Marco F. Huber; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4922-4932

Recent deep generative models (DGMs) such as generative adversarial networks (GANs) and diffusion probabilistic models (DPMs) have shown their impressive ability in generating high-fidelity photorealistic images. Although looking appealing to human eyes, training a model on purely synthetic images for downstream image processing tasks like image classification often results in an undesired performance drop compared to training on real data. Previous works have demonstrated that enhancing a real dataset with synthetic images from DGMs can be beneficial. However, the improvements were subjected to certain circumstances and yet were not comparable to adding the same number of real images. In this work, we propose a new taxonomy to describe factors contributing to this commonly observed phenomenon and investigate it on the popular CIFAR-10 dataset. We hypothesize that the Content Gap accounts for a large portion of the performance drop when using synthetic images from DGM and propose strategies to better utilize them in downstream tasks. Extensive experiments on multiple datasets showcase that our method outperforms baselines on downstream classification tasks both in case of training on synthetic only (Synthetic-to-Real) and training on a mix of real and synthetic data (Data Augmentation), particularly in the data-scarce scenario.

****

Learning Better Keypoints for Multi-Object 6DoF Pose Estimation

Yangzheng Wu, Michael Greenspan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 564-574

We address the problem of keypoint selection, and find that the performance of 6DoF pose estimation methods can be improved when pre-defined keypoint locations are learned, rather than being heuristically selected as has been the standard approach. We found that accuracy and efficiency can be improved by training a graph network to select a set of disperse keypoints with similarly distributed votes. These votes, learned by a regression network to accumulate evidence for the keypoint locations, can be regressed more accurately compared to previous heuristic keypoint algorithms. The proposed KeyGNet, supervised by a combined loss measuring both Wasserstein distance and dispersion, learns the color and geometry fe atures of the target objects to estimate optimal keypoint locations. Experiments demonstrate the keypoints selected by KeyGNet improved the accuracy for all evaluation metrics of all seven datasets tested, for three keypoint voting methods. The challenging Occlusion LINEMOD dataset notably improved ADD(S) by +16.4% on PVN3D, and all core BOP datasets showed an AR improvement for all objects, of between +1% and +21.5%. There was also a notable increase in performance when transitioning from single object to multiple object training using KeyGNet keypoints, essentially eliminating the SISO-MIMO gap for Occlusion LINEMOD.

****

Unsupervised Graphic Layout Grouping With Transformers

Jialiang Zhu, Danqing Huang, Chunyu Wang, Mingxi Cheng, Ji Li, Han Hu, Xin Geng, Baining Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1031-1040

Graphic design conveys messages through the combination of text, images and other visual elements. Unstructured designs such as overloaded social media graphics may fail to communicate their intended messages effectively. To address this issue, layout grouping offers a solution by organizing design elements into perceptual groups. While most methods rely on heuristic Gestalt principles, they often lack the context modeling ability needed to handle complex layouts. In this work, we reformulate the layout grouping task as a set prediction problem. It uses Transformers to learn a set of group tokens at various hierarchies, enabling it to reason the membership of the elements more effectively. The self-attention mechanism in Transformers boosts its context modeling ability, which enables it to handle complex layouts more accurately. To reduce annotation costs, we also propose an unsupervised learning strategy that pre-trains on noisy pseudo-labels induced by a novel heuristic algorithm. This approach then bootstraps to self-refine the noisy labels, further improving the accuracy of our model. Our extensive experiments demonstrate the effectiveness of our method, which outperforms existing state-of-the-art approaches in terms of accuracy and efficiency.

****

Can Vision-Language Models Be a Good Guesser? Exploring VLMs for Times and Location Reasoning

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, Volker Tresp; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 636-645

Vision-Language Models (VLMs) are expected to be capable of reasoning with commonsense knowledge as human beings. One example is that humans can reason where and when an image is taken based on their knowledge. This makes us wonder if, based on visual cues, Vision-Language Models that are pre-trained with large-scale image-text resources can achieve and even surpass human capability in reasoning times and location. To address this question, we propose a two-stage Recognition & Reasoning probing task applied to discriminative and generative VLMs to uncover whether VLMs can recognize times and location-relevant features and further reason about it. To facilitate the studies, we introduce WikiTiLo, a well-curated image dataset compromising images with rich socio-cultural cues. In extensive evaluation experiments, we find that although VLMs can effectively retain times and location-relevant features in visual encoders, they still fail to make perfect reasoning with context-conditioned visual features. The dataset is available at https://github.com/gengyuanmax/WikiTiLo.

****

What Decreases Editing Capability? Domain-Specific Hybrid Refinement for Improved GAN Inversion

Pu Cao, Lu Yang, Dongxv Liu, Xiaoya Yang, Tianrui Huang, Qing Song; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4240-4249

Recently, inversion methods have been exploring the incorporation of additional high-rate information from pretrained generators (such as weights or intermediate features) to improve the refinement of inversion and editing results from embedded latent codes. While such techniques have shown reasonable improvements in reconstruction, they often lead to a decrease in editing capability, especially w hen dealing with complex images that contain occlusions, detailed backgrounds, and artifacts. A vital crux is refining inversion results, avoiding editing capability degradation. To address this problem, we propose a novel refinement mechanism called Domain-Specific Hybrid Refinement (DHR), which draws on the advantages and disadvantages of two mainstream refinement techniques. We find that the weight modulation can gain favorable editing results but is vulnerable to these complex image areas and feature modulation is efficient at reconstructing. Hence, we divide the image into two domains and process them with these two methods separately. We first propose a Domain-Specific Segmentation module to automatically segment images into in-domain and out-of-domain parts according to their invertibility and editability without additional data annotation, where our hybrid refinement process aims to maintain the editing capability for in-domain areas and improve fidelity for both of them. We achieve this through Hybrid Modulation Refinement, which respectively refines these two domains by weight modulation and feature modulation. Our proposed method is compatible with all latent code embedding methods. Extension experiments demonstrate that our approach achieves state-of-the-art in real image inversion and editing. Code is available at https://github.com/caopulan/Domain-Specific_Hybrid_Refinement_Inversion.

****

Longformer: Longitudinal Transformer for Alzheimer's Disease Classification With Structural MRIs

Qiuhui Chen, Qiang Fu, Hao Bai, Yi Hong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3575-3584

Structural magnetic resonance imaging (sMRI), especially longitudinal sMRI, is often used to monitor and capture disease progression during the clinical diagnosis of Alzheimer's Disease (AD). However, current methods neglect AD's progressive nature and have mostly relied on a single image for recognizing AD. In this paper, we consider the problem of leveraging the longitudinal MRIs of a subject for AD classification. To address the challenges of missing data, data demand, and subtle changes over time in learning longitudinal 3D MRIs, we propose a novel model LongFormer, which is a hybrid 3D CNN and transformer design to learn from image and longitudinal flow pairs. Our model can fully leverage all images in a dataset and effectively fuse spatiotemporal features for classification. We evaluate our model on three datasets, i.e., ADNI, OASIS, and AIBL, and compare it to eight baseline algorithms. Our proposed LongFormer achieves state-of-the-art performance in classifying AD and NC subjects from all these three public datasets. Our source code is available online.

****

Grafting Vision Transformers

Jongwoo Park, Kumara Kahatapitiya, Donghyun Kim, Shivchander Sudalairaj, Quanfu Fan, Michael S. Ryoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1145-1154

Vision Transformers (ViTs) have recently become the state-of-the-art across many computer vision tasks. In contrast to convolutional networks (CNNs), ViTs enable global information sharing even within shallow layers of a network, i.e., among high-resolution features. However, this perk was later overlooked with the success of pyramid architectures such as Swin Transformer, which show better performance-complexity trade-offs. In this paper, we present a simple and efficient add-on component (termed GrafT) that considers global dependencies and multi-scale information throughout the network, in both high- and low-resolution features alike. It has the flexibility of branching out at arbitrary depths and shares most of the parameters and computations of the backbone. GrafT shows consistent gains over various well-known models which includes both hybrid and pure Transformer types, both homogeneous and pyramid structures, and various self-attention methods. In particular, it largely benefits mobile-size models by providing high-level semantics. On the ImageNet-1k dataset, GrafT delivers +3.9%, +1.4%, and +1.9% top-1 accuracy improvement to DeiT-T, Swin-T, and MobileViT-XXS, respectively. The code and models are at https://github.com/jongwoopark7978/Grafting-Vision-Transformer.

****

Hardware Aware Evolutionary Neural Architecture Search Using Representation Similarity Metric

Nilotpal Sinha, Abd El Rahman Shabayek, Anis Kacem, Peyman Rostami, Carl Shneider, Djamila Aouada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2628-2637

Hardware-aware Neural Architecture Search (HW-NAS) is a technique used to automatically design the architecture of a neural network for a specific task and target hardware. However, evaluating the performance of candidate architectures is a key challenge in HW-NAS, as it requires significant computational resources. To address this challenge, we propose an efficient hardware-aware evolution-based NAS approach called HW-EvRSNAS. Our approach re-frames the neural architecture search problem as finding an architecture with performance similar to that of a reference model for a target hardware, while adhering to a cost constraint for that hardware. This is achieved through a representation similarity metric known as Representation Mutual Information (RMI) employed as a proxy performance evaluator. It measures the mutual information between the hidden layer representations of a reference model and those of sampled architectures using a single training batch. We also use a penalty term that penalizes the search process in proportion to how far an architecture's hardware cost is from the desired hardware cost threshold. This resulted in a significantly reduced search time compared to the literature that reached up to 8000x speedups resulting in lower CO2 emissions. The proposed approach is evaluated on two different search spaces while using lower computational resources. Furthermore, our approach is thoroughly examined on six different edge devices under various hardware cost constraints.

****

DECDM: Document Enhancement Using Cycle-Consistent Diffusion Models

Jiaxin Zhang, Joy Rimchala, Lalla Mouatadid, Kamalika Das, Sricharan Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8036-8045

The performance of optical character recognition (OCR) heavily relies on document image quality, which is crucial for automatic document processing and document intelligence. However, most existing document enhancement methods require supervised data pairs, which raises concerns about data separation and privacy protection, and makes it challenging to adapt these methods to new domain pairs. To address these issues, we propose DECDM, an end-to-end document-level image translation method inspired by recent advances in diffusion models. Our method overcomes the limitations of paired training by independently training the source (noisy input) and target (clean output) models, making it possible to apply domain-specific diffusion models to other pairs. DECDM trains on one dataset at a time, eliminating the need to scan both datasets concurrently, and effectively preserving data privacy from the source or target domain. We also introduce simple data augmentation strategies to improve character-glyph conservation during translation. We compare DECDM with state-of-the-art methods on multiple synthetic data and benchmark datasets, such as document denoising and shadow removal, and demonstrate the superiority of performance quantitatively and qualitatively.

****

Watch Where You Head: A View-Biased Domain Gap in Gait Recognition and Unsupervised Adaptation

Gavriel Habib, Noa Barzilay, Or Shimshi, Rami Ben-Ari, Nir Darshan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6109-6119

Gait Recognition is a computer vision task aiming to identify people by their walking patterns. Although existing methods often show high performance on specific datasets, they lack the ability to generalize to unseen scenarios. Unsupervised Domain Adaptation (UDA) tries to adapt a model, pre-trained in a supervised manner on a source domain, to an unlabelled target domain. There are only a few works on UDA for gait recognition proposing solutions to limited scenarios. In this paper, we reveal a fundamental phenomenon in adaptation of gait recognition models, caused by the bias in the target domain to viewing angle or walking direction. We then suggest a remedy to reduce this bias with a novel triplet selection strategy combined with curriculum learning. To this end, we present Gait Orientation-based method for Unsupervised Domain Adaptation (GOUDA). We provide extensive experiments on four widely-used gait datasets, CASIA-B, OU-MVLP, GREW, and Gait3D, and on three backbones, GaitSet, GaitPart, and GaitGL, justifying the view bias and showing the superiority of our proposed method over prior UDA works.

****

Show Your Face: Restoring Complete Facial Images From Partial Observations for VR Meeting

Zheng Chen, Zhiqi Zhang, Junsong Yuan, Yi Xu, Lantao Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8688-8697

Virtual Reality (VR) headsets allow users to interact with the virtual world. However, the device physically blocks visual connections among users, causing huge inconveniences for VR meetings. To address this issue, studies have been conducted to restore human faces from images captured by Headset Mounted Cameras (HMC). Unfortunately, existing approaches heavily rely on high-resolution person-specific 3D models which are prohibitively expensive to apply to large-scale scenarios. Our goal is to design an efficient framework for restoring users' facial data in VR meetings. Specifically, we first build a new dataset, named Facial Image Composition (FIC) data which approximates the real HMC images from a VR headset. By leveraging the heterogeneity of the HMC images, we decompose the restoration problem into a local geometry transformation and global color/style fusion. Then we propose a 2D light-weight facial image composition network (FIC-Net), where three independent local models are responsible for transforming raw HMC patches and the global model performs a fusion of the transformed HMC patches with a pre-recorded reference image. Finally, we also propose a stage-wise training strategy to optimize the generalization of our FIC-Net. We have validated the effectiveness of our proposed FIC-Net through extensive experiments.

****

Shape From Shading for Robotic Manipulation

Arkadeep Narayan Chaudhury, Leonid Keselman, Christopher G. Atkeson; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8389-8398

Controlling illumination can generate high quality information about object surface normals and depth discontinuities at a low computational cost. In this work we demonstrate a robot workspace-scaled controlled illumination approach that generates high quality information for table top scale objects for robotic manipulation. With our low angle of incidence directional illumination approach, we can precisely capture surface normals and depth discontinuities of monochromatic Lambertian objects. We show that this approach to shape estimation is 1) valuable for general purpose grasping with a single point vacuum gripper, 2) can measure the deformation of known objects, and 3) can estimate pose of known objects and track unknown objects in the robot's workspace.

****

Self-Supervised Denoising Transformer With Gaussian Process

Rajeev Yasarla, Jeya Maria Jose Valanarasu, Vishwanath Sindagi, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1474-1484

Convolutional neural network (CNN) based methods have been the main focus of recent developments for image denoising. However, these methods lack majorly in two ways: 1) They require a large amount of labeled data to perform well. 2) They do not have a good global understanding due to convolutional inductive biases. Recent emergence of Transformers and self-supervised learning methods have focused on tackling these issues. In this work, we address both these issues for image denoising and propose a new method: Self-Supervised denoising Transformer (SST-GP) with Gaussian Process. Our novelties are two fold: First, we propose a new way of doing self-supervision by incorporating Gaussian Processes (GP). Given a noisy image, we generate multiple noisy down-sampled images with random cyclic shifts. Using GP, we formulate a joint Gaussian distribution between these down-sampled images and learn the relation between their corresponding denoising functio

n mappings to predict the pseudo-Ground truth (pseudo-GT) for each of the down-sampled images. This enables the network to learn noise present in the down-sampled images and achieve better denoising performance by using the joint relationship between down-sampled images with help of GP. Second, we propose a new transformer architecture - Denoising Transformer (Den-T) which is tailor-made for denoising application. Den-T has two transformer encoder branches - one which focuses on extracting fine context details and another to extract coarse context details. This helps Den-T to attend to both local and global information to effectively denoise the image. Finally, we train Den-T using the proposed self-supervised strategy using GP and achieve a better performance over recent unsupervised/self-supervised denoising approaches when validated on various denoising datasets like Kodak, BSD, Set-14 and SIDD. Codes will be made public after review.
****

SemST: Semantically Consistent Multi-Scale Image Translation via Structure-Texture Alignment

Ganning Zhao, Wenhui Cui, Suya You, C.-C. Jay Kuo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7209-7219

Unsupervised image-to-image translation learns cross-domain image mapping that transfers input from the source domain to output in the target domain while preserving its semantics. One challenge is that different semantic statistics in source and target domains result in content discrepancy known as semantic distortion. To address this problem, a novel I2I method that maintains semantic consistency in translation is proposed and named SemST in this work. SemST reduces semantic distortion by employing contrastive learning and aligning the structural and textural properties of input and output by maximizing their mutual information. Furthermore, a multi-scale approach is introduced to enhance translation performance, thereby enabling the applicability of SemST to domain adaptation in high-resolution images. Experiments show that SemST effectively mitigates semantic distortion and achieves state-of-the-art performance. Also, the application of SemST to domain adaptation is explored. It is demonstrated by preliminary experiments that SemST can be utilized as a beneficial pre-training for the semantic segmentation task.
****

Bag of Tricks for Fully Test-Time Adaptation

Saypraseuth Mounsaveng, Florent Chiaroni, Malik Boudiaf, Marco Pedersoli, Ismail Ben Ayed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1936-1945

Fully Test-Time Adaptation (TTA), which aims at adapting models to data drifts, has recently attracted wide interest. Numerous tricks and techniques have been proposed to ensure robust learning on arbitrary streams of unlabeled data. However, assessing the true impact of each individual technique and obtaining a fair comparison still constitutes a significant challenge. To help consolidate the community's knowledge, we present a categorization of selected orthogonal TTA techniques, including small batch normalization, stream rebalancing, reliable sample selection, and network confidence calibration. We meticulously dissect the effect of each approach on different scenarios of interest. Through our analysis, we shed light on trade-offs induced by those techniques between accuracy, the computational power required, and model complexity. We also uncover the synergy that arises when combining techniques and are able to establish new state-of-the-art results.
****

OE-CTST: Outlier-Embedded Cross Temporal Scale Transformer for Weakly-Supervised Video Anomaly Detection

Snehashis Majhi, Rui Dai, Quan Kong, Lorenzo Garattoni, Gianpiero Francesca, François Brémond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8574-8583

Video anomaly detection in real-world scenarios is challenging due to the complex temporal blending of long and short-length anomalies with normal ones. Further, it is more difficult to detect those due to : (i) Distinctive features characterizing the short and long anomalies with sharp and progressive temporal cues re spectively; (ii) Lack of precise temporal information (i.e. weak-supervision) limits the temporal dynamics modeling of anomalies from normal events. In this paper, we propose a novel 'temporal transformer' framework for weakly-supervised anomaly detection: OE-CTST. The proposed framework has two major components: (i) Outlier Embedder (OE) and (ii) Cross Temporal Scale Transformer (CTST). First, OE generates anomaly-aware temporal position encoding to allow the transformer to effectively model the temporal dynamics among the anomalies and normal events. Second, CTST encodes the cross-correlation between multi-temporal scale features to benefit short and long length anomalies by modeling the global temporal relations. The proposed OE-CTST is validated on three publicly available datasets i.e. UCF-Crime, XD-Violence, and IITB-Corridor, outperforming recently reported state-of-the-art approaches.

****

Bridging Generalization Gaps in High Content Imaging Through Online Self-Supervised Domain Adaptation

Johan Fredin Haslum, Christos Matsoukas, Karl-Johan Leuchowius, Kevin Smith; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7738-7747

High Content Imaging (HCI) plays a vital role in modern drug discovery and development pipelines, facilitating various stages from hit identification to candidate drug characterization. Applying machine learning models to these datasets can prove challenging as they typically consist of multiple batches, affected by experimental variation, especially if different imaging equipment have been used. Moreover, as new data arrive, it is preferable that they are analyzed in an online fashion. To overcome this, we propose CODA, an online self-supervised domain adaptation approach. CODA divides the classifier's role into a generic feature extractor and a task-specific model. We adapt the feature extractor's weights to the new domain using cross-batch self-supervision while keeping the task-specific model unchanged. Our results demonstrate that this strategy significantly reduces the generalization gap, achieving up to a 300% improvement when applied to data from different labs utilizing different microscopes. CODA can be applied to new, unlabeled out-of-domain data sources of different sizes, from a single plate to multiple experimental batches.

****

Using Early Readouts To Mediate Featural Bias in Distillation

Rishabh Tiwari, Durga Sivasubramanian, Anmol Mekala, Ganesh Ramakrishnan, Pradeep Shenoy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2638-2647

Deep networks tend to learn spurious feature-label correlations in real-world supervised learning tasks. This vulnerability is aggravated in distillation, where a student model may have lesser representational capacity than the corresponding teacher model. Often, knowledge of specific spurious correlations is used to reweight instances & rebalance the learning process. We propose a novel early readout mechanism whereby we attempt to predict the label using representations from earlier network layers. We show that these early readouts automatically identify problem instances or groups in the form of confident, incorrect predictions. Leveraging these signals to modulate the distillation loss on an instance level allows us to substantially improve not only group fairness measures across benchmark datasets, but also overall accuracy of the student model. We also provide secondary analyses that bring insight into the role of feature learning in supervision and distillation.

****

Continuous Adaptation for Interactive Segmentation Using Teacher-Student Architecture

Barsegh Atanyan, Levon Khachatryan, Shant Navasardyan, Yunchao Wei, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 789-799

Interactive segmentation is the task of segmenting objects or regions of interest from images based on user annotations. While most current methods perform effectively on images from the same distribution as the training dataset, they suffe r to generalize on unseen domains. To address this issue some approaches incorporate test-time adaptation techniques which, on the other hand, may lead to catastrophic forgetting (i.e. degrading the performance on the previously seen domains) when applied on datasets from various domains sequentially.In this paper, we propose a novel domain adaptation approach leveraging a teacher-student learning framework to tackle the catastrophic forgetting issue. Continuously updating the student and teacher models based on user clicks results in improved segmentation accuracy on unseen domains, while preserving comparable performance on previous domains.Our approach is evaluated on a sequence of datasets from unseen domains (i.e. medical, aerial images, etc.), and, after adaptation, on the source domain demonstrating a significant decline of catastrophic forgetting (e.g. from 55% to 4% on Berkeley dataset).

****

Causal Feature Alignment: Learning To Ignore Spurious Background Features

Rahul Venkataramani, Parag Dutta, Vikram Melapudi, Ambedkar Dukkipati; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4666-4674

Deep neural networks are susceptible to spurious features strongly correlating with the target. This phenomenon leads to sub-optimal performance during real-world deployment where the spurious correlations do not exist, leading to deployment challenges in safety-critical environments like healthcare, autonomous navigation etc. While spurious features can correlate with causal features in myriad ways, we propose a solution for a common manifestation in computer vision where the background corresponds to a spurious feature. In contrast to previous works, we do not require apriori knowledge of different sub-groups in the data induced by the presence/absence of spurious features and the corresponding access to samples from these sub-groups. Our proposed method, Causal Feature Alignment (CFA), utilizes segmentation of foreground (a proxy for the causal component) on a small subset of training examples to align the representations of the original images to match words from only causal elements. We first demonstrate the validity of the proposed method on semi-synthetic data. Subsequently, we obtain state-of-the-art results on worst-group accuracy (93%) on the benchmark dataset of Waterbirds using CFA. Furthermore, we demonstrate significant gains of 6% on the Backgrounds Challenge. Finally, we show that utilizing the recently released foundational methods can alleviate the requirement of dense segmentation and can be substituted with weaker modes of human input like bounding boxes, clicks etc., without any performance loss compared to the original CFA.

****

VD-GR: Boosting Visual Dialog With Cascaded Spatial-Temporal Multi-Modal Graphs

Adnen Abdessaied, Lei Shi, Andreas Bulling; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5805-5814

We propose VD-GR -- a novel visual dialog model that combines pre-trained language models (LMs) with graph neural networks (GNNs). Prior works mainly focused on one class of models at the expense of the other, thus missing out on the opportunity of combining their respective benefits. At the core of VD-GR is a novel integration mechanism that alternates between spatial-temporal multi-modal GNNs and BERT layers, and that covers three distinct contributions: First, we use multi-modal GNNs to process the features of each modality (image, question, and dialog history) and exploit their local structures before performing BERT global attention. Second, we propose hub-nodes that link to all other nodes within one modality graph, allowing the model to propagate information from one GNN (modality) to the other in a cascaded manner. Third, we augment the BERT hidden states with fine-grained multi-modal GNN features before passing them to the next VD-GR layer. Evaluations on VisDial v1.0, VisDial v0.9, VisDialConv, and VisPro show that VD-GR achieves new state-of-the-art results on all datasets.

****

Fingervein Verification Using Convolutional Multi-Head Attention Network

Raghavendra Ramachandra, Sushma Venkatesh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6175-6184

Biometric verification systems are deployed in various security-based access-con

trol applications that require user-friendly and reliable user verification. Among the different biometric characteristics, fingervein biometrics have been extensively studied owing to their reliable verification performance. Furthermore, fingervein patterns reside inside the skin and are not visible outside; therefore, they possess inherent resistance to presentation attacks and degradation due to external factors. In this study, we introduce a novel fingervein verification technique using a convolutional multihead attention network, VeinAtnNet. The proposed VeinAtnNet is designed to achieve light weight with a smaller number of learnable parameters while extracting discriminant information from both normal and enhanced fingervein images. The proposed VeinAtnNet was trained on the newly constructed fingervein dataset with 300 unique fingervein patterns that were captured in multiple sessions to obtain 92 samples per unique fingervein. Extensive experiments were performed on the newly collected dataset FV-300 and the publicly available FV-USM fingervein dataset. The performance of the proposed method was compared with five state-of-the-art fingervein verification systems, indicating the efficacy of the proposed VeinAtnNet.
****

Foundation Model Assisted Weakly Supervised Semantic Segmentation
Xiaobo Yang, Xiaojin Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 523-532

This work aims to leverage pre-trained foundation models, such as contrastive language-image pre-training (CLIP) and segment anything model (SAM), to address weakly supervised semantic segmentation (WSSS) using image-level labels. To this end, we propose a coarse-to-fine framework based on CLIP and SAM for generating high-quality segmentation seeds. Specifically, we construct an image classification task and a seed segmentation task, which are jointly performed by CLIP with frozen weights and two sets of learnable task-specific prompts. A SAM-based seeding (SAMS) module is designed and applied to each task to produce either coarse or fine seed maps. Moreover, we design a multi-label contrastive loss supervised by image-level labels and a CAM activation loss supervised by the generated coarse seed map. These losses are used to learn the prompts, which are the only parts need to be learned in our framework. Once the prompts are learned, we input each image along with the learned segmentation-specific prompts into CLIP and the SAMS module to produce high-quality segmentation seeds. These seeds serve as pseudo labels to train an off-the-shelf segmentation network like other two-stage WSSS methods. Experiments show that our method achieves the state-of-the-art performance on PASCAL VOC 2012 and competitive results on MS COCO 2014. Our code will be released upon acceptance.
****

Describe Images in a Boring Way: Towards Cross-Modal Sarcasm Generation
Jie Ruan, Yue Wu, Xiaojun Wan, Yuesheng Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5701-5710

Sarcasm generation has been investigated in previous studies by considering it as a text-to-text generation problem, i.e., generating a sarcastic sentence for an input sentence. In this paper, we study a new problem of cross-modal sarcasm generation (CMSG), i.e., generating a sarcastic description for a given image. CMSG is challenging as models need to satisfy the characteristics of sarcasm, as well as the correlation between different modalities. In addition, there should be some inconsistency between the two modalities, which requires imagination. Moreover, high-quality training data is insufficient. To address these problems, we take a step toward generating sarcastic descriptions from images without paired training data and propose an Extraction-Generation-Ranking based Modular method (EGRM) for CMSG. Specifically, EGRM first extracts diverse information from an image at different levels and uses the obtained image tags, sentimental descriptive caption, and commonsense-based consequence to generate candidate sarcastic texts. Then, a comprehensive ranking algorithm, which considers image-text relation, sarcasticness, and grammaticality, is proposed to select a final text from the candidate texts. Human evaluation at five criteria on a total of 2100 generated image-text pairs and auxiliary automatic evaluation show the superiority of our method. Code and data will be publicly available.

****

Offline-to-Online Knowledge Distillation for Video Instance Segmentation

Hojin Kim, Seunghun Lee, Hyeon Kang, Sunghoon Im; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 159-168

In this paper, we present offline-to-online knowledge distillation (OOKD) for video instance segmentation (VIS), which transfers a wealth of video knowledge from an offline model to an online model for consistent prediction. Unlike previous methods that have adopted either an online or offline model, our single online model takes advantage of both models by distilling offline knowledge. To transfer knowledge correctly, we propose query filtering and association (QFA), which filters irrelevant queries to exact instances. Our KD with QFA increases the robustness of feature matching by encoding object-centric features from a single frame supplemented by long-range global information. We also propose a simple data augmentation scheme for knowledge distillation in the VIS task that fairly transfers the knowledge of all classes into the online model. Extensive experiments show that our method significantly improves the performance in video instance segmentation, especially for challenging datasets, including long, dynamic sequences. Our method also achieves state-of-the-art performance on YTVIS-21, YTVIS-22, and OVIS datasets, with mAP scores of 46.1%, 43.6%, and 31.1%, respectively.

****

Rethinking Multimodal Content Moderation From an Asymmetric Angle With Mixed-Modality

Jialin Yuan, Ye Yu, Gaurav Mittal, Matthew Hall, Sandra Sajeev, Mei Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8532-8542

There is a rapidly growing need for multimodal content moderation (CM) as more and more content on social media is multimodal in nature. Existing unimodal CM systems may fail to catch harmful content that crosses modalities (e.g., memes or videos), which may lead to severe consequences. In this paper, we present a novel CM model, Asymmetric Mixed-Modal Moderation (AM3), to target multimodal and unimodal CM tasks. Specifically, to address the asymmetry in semantics between vision and language, AM3 has a novel asymmetric fusion architecture that is designed to not only fuse the common knowledge in both modalities but also to exploit the unique information in each modality. Unlike pre- vious works that focus on representing the two modalities in similar feature space while overlooking the intrinsic difference between the information conveyed in multimodality and in unimodality (asymmetry in modalities), we propose a novel cross-modality contrastive loss to learn the unique knowledge that only appears in multimodality. This is critical as some harmful intent may only be conveyed through the intersection of both modalities. With extensive experiments, we show that AM3 outperforms all existing state-of-the-art methods on both multimodal and unimodal CM benchmarks.

****

Active Learning With Task Consistency and Diversity in Multi-Task Networks

Aral Hekimoglu, Michael Schmidt, Alvaro Marcos-Ramiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2503-2512

Multi-task networks demonstrate state-of-the-art performance across various vision tasks. However, their performance relies on large-scale annotated datasets, demanding extensive labeling efforts, especially as the number of tasks to label increases. In this paper, we introduce an active learning framework consisting of a data selection strategy that identifies the most informative unlabeled samples and a training strategy that ensures balanced training across multiple tasks. Our selection strategy leverages the inconsistency between initial and refined task predictions generated by recent two-stage multi-task networks. We further enhance our selection by incorporating task-specific sample diversity through a novel feature extraction mechanism. Our method captures task features for all tasks and distills them into a unified representation, which is used to curate a training set encapsulating diverse task-specific scenarios. In our training strategy, we introduce a sample-specific loss weighting mechanism based on the individual task selection scores. This facilitates the individual prioritization of sam ples for each task, effectively simulating the sample ordering process inherent in single-task active learning. Extensive experimentation on the PASCAL and NYUD-v2 datasets demonstrates that our approach outperforms existing state-of-the-art methods. Our approach reaches the loss of the network trained with all the available data using only 50% of the data, corresponding to 10% fewer labels compared to the state-of-the-art selection strategy. Our code is available at https://github.com/aralhekimoglu/mtal.

****

Single Domain Generalization via Normalised Cross-Correlation Based Convolutions

WeiQin Chuah, Ruwan Tennakoon, Reza Hoseinnezhad, David Suter, Alireza Bab-Hadiashar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1752-1761

Deep learning techniques often perform poorly in the presence of domain shift, where the test data follows a different distribution than the training data. The most practically desirable approach to address this issue is Single Domain Generalization (S-DG), which aims to train robust models using data from a single source. Prior work on S-DG has primarily focused on using data augmentation techniques to generate diverse training data. In this paper, we explore an alternative approach by investigating the robustness of linear operators, such as convolution and dense layers commonly used in deep learning. We propose a novel operator called XCNorm that computes the normalized cross-correlation between weights and an input feature patch. This approach is invariant to both affine shifts and changes in energy within a local feature patch and eliminates the need for commonly used non-linear activation functions. We show that deep neural networks composed of this operator are robust to common semantic distribution shifts. Furthermore, our empirical results on single-domain generalization benchmarks demonstrate that our proposed technique performs comparably to the state-of-the-art methods.

****

Intrinsic Hand Avatar: Illumination-Aware Hand Appearance and Shape Reconstruction From Monocular RGB Video

Pratik Kalshetti, Parag Chaudhuri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6120-6130

Reconstructing a user-specific hand avatar is essential for a personalized experience in augmented and virtual reality systems. Current state-of-the-art avatar reconstruction methods use implicit representations to capture detailed geometry and appearance combined with neural rendering. However, these methods rely on a complicated multi-view setup, do not explicitly handle environment lighting leading to baked-in illumination and self-shadows, and require long hours for training. We present a method to reconstruct a hand avatar from a monocular RGB video of a user's hand in arbitrary hand poses captured under real-world environment lighting. Specifically, our method jointly optimizes shape, appearance, and lighting parameters using a realistic shading model in a differentiable rendering framework incorporating Monte Carlo path tracing. Despite relying on physically-based rendering, our method can complete the reconstruction within minutes. In contrast to existing work, our method disentangles intrinsic properties of the underlying appearance and environment lighting, leading to realistic self-shadows. We compare our method with state-of-the-art hand avatar reconstruction methods and observe that it outperforms them on all commonly used metrics. We also evaluate our method on our captured dataset to emphasize its generalization capability. Finally, we demonstrate applications of our intrinsic hand avatar on novel pose synthesis and relighting. We plan to release our code to aid further research.

****

Object Re-Identification From Point Clouds

Benjamin Thérien, Chengjie Huang, Adrian Chow, Krzysztof Czarnecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8377-8388

Object re-identification (ReID) from images plays a critical role in application domains of image retrieval (surveillance, retail analytics, etc.) and multi-object tracking (autonomous driving, robotics, etc.). However, systems that additionally or exclusively perceive the world from depth sensors are becoming more com

monplace without any corresponding methods for object ReID. In this work, we fill the gap by providing the first large-scale study of object ReID from point clouds and establishing its performance relative to image ReID. To enable such a study, we create two large-scale ReID datasets with paired image and LiDAR observations and propose a lightweight matching head that can be concatenated to any set or sequence processing backbone (e.g., PointNet or ViT), creating a family of comparable object ReID networks for both modalities. Run in Siamese style, our proposed point cloud ReID networks can make thousands of pairwise comparisons in real-time (10 hz). Our findings demonstrate that their performance increases with higher sensor resolution and approaches that of image ReID when observations are sufficiently dense. Our strongest network trained at the largest scale achieves ReID accuracy exceeding 90% for rigid objects and 85% for deformable objects (without any explicit skeleton normalization). To our knowledge, we are the first to study object re-identification from real point cloud observations.
****

MotionGPT: Human Motion Synthesis With Improved Diversity and Realism via GPT-3 Prompting

Jose Ribeiro-Gomes, Tianhui Cai, Zoltán Á. Milacski, Chen Wu, Aayush Prakash, Shingo Takagi, Amaury Aubel, Daeil Kim, Alexandre Bernardino, Fernando De la Torre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5070-5080

There are numerous applications for human motion synthesis, including animation, gaming, robotics, or sports science. In recent years, human motion generation from natural language has emerged as a promising alternative to costly and labor-intensive data collection methods relying on motion capture or wearable sensors (e.g., suits). Despite this, generating human motion from textual descriptions remains a challenging and intricate task, primarily due to the scarcity of large-scale supervised datasets capable of capturing the full diversity of human activity. This study proposes a new approach, called MotionGPT, to address the limitations of previous text-based human motion generation methods by utilizing the extensive semantic information available in large language models (LLMs). We first pretrain a doubly text-conditional motion diffusion model on both coarse ("high-level") and detailed ("low-level") ground truth text data. Then during inference, we improve motion diversity and alignment with the training set, by zero-shot prompting GPT-3 for additional "low-level" details. Our method achieves new state-of-the-art quantitative results in terms of Frechet Inception Distance (FID) and motion diversity metrics, and improves all considered metrics. Furthermore, it has strong qualitative performance, producing natural results.
****

Training-Based Model Refinement and Representation Disagreement for Semi-Supervised Object Detection

Seyed Mojtaba Marvasti-Zadeh, Nilanjan Ray, Nadir Erbilgin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2256-2265

Semi-supervised object detection (SSOD) aims to improve the performance and generalization of existing object detectors by utilizing limited labeled data and extensive unlabeled data. Despite many advances, recent SSOD methods are still challenged by inadequate model refinement using the classical exponential moving average (EMA) strategy, the consensus of Teacher-Student models in the latter stages of training (i.e., losing their distinctiveness), and noisy/misleading pseudo-labels. This paper proposes a novel training-based model refinement (TMR) stage and a simple yet effective representation disagreement (RD) strategy to address the limitations of classical EMA and the consensus problem. The TMR stage of Teacher-Student models optimizes the lightweight scaling operation to refine the model's weights and prevent overfitting or forgetting learned patterns from unlabeled data. Meanwhile, the RD strategy helps keep these models diverged to encourage the student model to explore additional patterns in unlabeled data. Our approach can be integrated into established SSOD methods and is empirically validated using two baseline methods, with and without cascade regression, to generate more reliable pseudo-labels. Extensive experiments demonstrate the superior perfo rmance of our approach over state-of-the-art SSOD methods. Specifically, the proposed approach outperforms the baseline Unbiased-Teacher-v2 (& Unbiased-Teacher-v1) method by an average mAP margin of 2.23, 2.1, and 3.36 (& 2.07, 1.9, and 3.27) on COCO-standard, COCO-additional, and Pascal VOC datasets, respectively.

****

Efficient Layout-Guided Image Inpainting for Mobile Use

Wenbo Li, Yi Wei, Yilin Shen, Hongxia Jin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8450-8459

The layout guidance, which specifies the pixel-wise object distribution, is beneficial to preserving the object boundaries in image inpainting while not hurting model's generalization capability. We aim to design an efficient and robust layout-guided image inpainting method for mobile use, which can achieve the robustness in presence of the mixed scenes where objects with the delicate shape reside next to the hole. Our method is made up of two sub-models, which restore the pixel-information for the hole from coarse to fine, and support each other to overcome the practical challenges encountered when making the whole method lightweight. The layout mask guides the two sub-models, which thus enables the robustness of our method in mixed scenes. We demonstrate the efficiency and robustness of our method via both the experiments and a mobile demo.

****

SigmML: Metric Meta-Learning for Writer Independent Offline Signature Verification in the Space of SPD Matrices

Alexios Giazitzis, Elias N. Zois; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6312-6322

The handwritten signature has been identified as one of the most popular biometric means of human consent and/or presence for transactions held by any number of physical or legal entities. Automated signature verification (ASV), merge popular scientific branches such as computer vision, pattern recognition and/or data-driven machine learning algorithms. Up to now, several metric learning approaches for designing a writer-independent signature verifier, have been developed within a Euclidean framework by means of having their operations closed with respect to real vector spaces. In this work, we propose, for the first time in the ASV literature, the use of a meta-learning framework in the space of the Symmetric Positive Definite (SPD) manifold as a means to learn a pairwise similarity metric for writer-independent ASV. To begin, pairs of handwritten signatures are converted into a multidimensional distance vector with elements corresponding SPD distances between spatial segments of corresponding covariance pairs. We propose a novel meta-learning approach which explores the structure of the input gradients of the SPD manifold by means of a recurrent model, constrained by the geometry of the SPD manifold. The experimental protocols utilize two popular signature datasets of Western and Asian origin in two blind-intra and blind-inter (or cross-lingual) transfer learning approach. It also provide evidence of the discriminating nature of the proposed framework at least when summarized against other State-of-the-Art models, typically realized under a framework of Euclidean, or vector space, nature.

****

Mini but Mighty: Finetuning ViTs With Mini Adapters

Imad Eddine Marouf, Enzo Tartaglione, Stéphane Lathuilière; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1732-1741

Vision Transformers (ViTs) have become one of the dominant architectures in computer vision, and pre-trained ViT models are commonly adapted to new tasks via fine-tuning. Recent works proposed several parameter-efficient transfer learning methods, such as adapters, to avoid the prohibitive training and storage cost of fine-tuning. In this work, we observe that adapters perform poorly when the dimension of adapters is small, and we propose MiMi, a training framework that addresses this issue. We start with large adapters which can reach high performance, and iteratively reduce the size of every adapter. We introduce a scoring function that compares neuron importance across layers and consequently allows automatic estimation of the hidden dimension of every adapter. Our method outperforms ex isting methods in finding the best trade-off between accuracy and trained parameters across the three dataset benchmarks DomainNet, VTAB, and Multi-task, for a total of 29 datasets. We will release our code publicly upon acceptance.

****

Dynamic Multimodal Information Bottleneck for Multimodality Classification

Yingying Fang, Shuang Wu, Sheng Zhang, Chaoyan Huang, Tieyong Zeng, Xiaodan Xing, Simon Walsh, Guang Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7696-7706

Effectively leveraging multimodal data such as various images, laboratory tests and clinical information is gaining traction in a variety of AI-based medical diagnosis and prognosis tasks. Most existing multi-modal techniques only focus on enhancing their performance by leveraging the differences or shared features from various modalities and fusing feature across different modalities. These approaches are generally not optimal for clinical settings, which pose the additional challenges of limited training data, as well as being rife with redundant data or noisy modality channels, leading to subpar performance. To address this gap, we study the robustness of existing methods to data redundancy and noise and propose a generalized dynamic multimodal information bottleneck framework for attaining a robust fused feature representation. Specifically, our information bottleneck module serves to filter out the task-irrelevant information and noises in the fused feature, and we further introduce a sufficiency loss to prevent dropping of task-relevant information, thus explicitly preserving the sufficiency of prediction information in the distilled feature. We validate our model on an in-house and a public COVID-19 dataset for mortality prediction as well as two public biomedical datasets for diagnostic tasks. Extensive experiments show that our method surpasses the state-of-the-art and is significantly more robust, being the only method to remain performant when large-scale noisy channels exist. Our code is publicly available at https://github.com/Anonymous-PaperSubmission/DMIB.

****

Learning Generalizable Perceptual Representations for Data-Efficient No-Reference Image Quality Assessment

Suhas Srinath, Shankhanil Mitra, Shika Rao, Rajiv Soundararajan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 22-31

No-reference (NR) image quality assessment (IQA) is an important tool in enhancing the user experience in diverse visual applications. A major drawback of state-of-the-art NR-IQA techniques is their reliance on a large number of human annotations to train models for a target IQA application. To mitigate this requirement, there is a need for unsupervised learning of generalizable quality representations that capture diverse distortions. We enable the learning of low-level quality features agnostic to distortion types by introducing a novel quality-aware contrastive loss. Further, we leverage the generalizability of vision-language models by fine-tuning one such model to extract high-level image quality information through relevant text prompts. The two sets of features are combined to effectively predict quality by training a simple regressor with very few samples on a target dataset. Additionally, we design zero-shot quality predictions from both pathways in a completely blind setting. Our experiments on diverse datasets encompassing various distortions show the generalizability of the features and their superior performance in the data-efficient and zero-shot settings.

****

Real Time GAZED: Online Shot Selection and Editing of Virtual Cameras From Wide-Angle Monocular Video Recordings

Sudheer Achary, Rohit Girmaji, Adhiraj Anil Deshmukh, Vineet Gandhi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4108-4116

Eliminating time-consuming post-production processes and delivering high-quality videos in today's fast-paced digital landscape are the key advantages of real-time approaches. To address these needs, we present Real Time GAZED: a real-time adaptation of the GAZED framework integrated with CineFilter, a novel real-time camera trajectory stabilization approach. It enables users to create professiona

lly edited videos in real-time. Comparative evaluations against baseline methods, including the non-real-time GAZED, demonstrate that Real Time GAZED achieves similar editing results, ensuring high-quality video output. Furthermore, a user study confirms the aesthetic quality of the video edits produced by the Real Time GAZED approach. With these advancements in real-time camera trajectory optimization and video editing presented, the demand for immediate and dynamic content creation in industries such as live broadcasting, sports coverage, news reporting, and social media content creation can be met more efficiently.
****

ConfTrack: Kalman Filter-Based Multi-Person Tracking by Utilizing Confidence Score of Detection Box

Hyeonchul Jung, Seokjun Kang, Takgen Kim, HyeongKi Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6583-6592

Kalman filter-based tracking-by-detection (KFTBD) trackers are effective methods for solving multi-person tracking tasks. However, in crowd circumstances, noisy detection results (bounding boxes with low-confidence scores) can cause ID switch and tracking failure of trackers since these trackers utilize the detector's output directly. In this paper, to solve the problem, we suggest a novel tracker called ConfTrack based on a KFTBD tracker. Compared with conventional KFTBD trackers, ConfTrack consists of novel algorithms, including low-confidence object penalization and cascading algorithms for effectively dealing with noisy detector outputs. ConfTrack is tested on diverse domains of datasets such as the MOT17, MOT20, DanceTrack, and HiEve datasets. ConfTrack has proved its robustness in crowd circumstances by achieving the highest score at HOTA and IDF1 metrics in the MOT20 dataset.
****

Hybrid Sample Synthesis-Based Debiasing of Classifier in Limited Data Setting

Piyush Arora, Pratik Mazumder; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4791-4799

Deep learning models are known to suffer from the problem of bias, and researchers have been exploring methods to address this issue. However, most of these methods require prior knowledge of the bias and are not always practical. In this paper, we focus on a more practical setting with no prior information about the bias. Generally, in this setting, there are a large number of bias-aligned samples that cause the model to produce biased predictions and a few bias-conflicting samples that do not conform to the bias. If the training data is limited, the influence of the bias-aligned samples may become even stronger on the model predictions, and we experimentally demonstrate that existing debiasing techniques suffer severely in such cases. In this paper, we examine the effects of unknown bias in small dataset regimes and present a novel approach to mitigate this issue. The proposed approach directly addresses the issue of the extremely low occurrence of bias-conflicting samples in limited data settings through the synthesis of hybrid samples that can be used to reduce the effect of bias. We perform extensive experiments on several benchmark datasets and experimentally demonstrate the effectiveness of our proposed approach in addressing any unknown bias in the presence of limited data. Specifically, our approach outperforms the vanilla, LfF, LDD, and DebiAN debiasing methods by absolute margins of 10.39%, 9.08%, 8.07%, and 9.67% when only 10% of the Corrupted CIFAR-10 Type 1 dataset is available with a bias-conflicting sample ratio of 0.05.
****

Visually Guided Audio Source Separation With Meta Consistency Learning

Md Amirul Islam, Seyed Shahabeddin Nabavi, Irina Kezele, Yang Wang, Yuanhao Yu, Jin Tang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3014-3023

In this paper, we tackle the problem of visually guided audio source separation in the context of both known and unknown objects (e.g., musical instruments). Recent successful end-to-end deep learning approaches adopt a single network with fixed parameters to generalize across unseen test videos. However, it can be challenging to generalize in cases where the distribution shift between training an d test videos is higher as they fail to utilize internal information of unknown test videos. Based on this observation, we introduce a meta-consistency driven test time adaptation scheme that enables the pretrained model to quickly adapt to known and unknown test music videos in order to bring substantial improvements. In particular, we design a self-supervised audio-visual consistency objective as an auxiliary task that learns the synchronization between audio and its corresponding visual embedding. Concretely, we apply a meta-consistency training scheme to further optimize the pretrained model for effective and faster test time adaptation. We obtain substantial performance gains with only a smaller number of gradient updates and without any additional parameters for the task of audio source separation. Extensive experimental results across datasets demonstrate the effectiveness of our proposed method.

****

RGBT-Dog: A Parametric Model and Pose Prior for Canine Body Analysis Data Creation

Jake Deane, Sinéad Kearney, Kwang In Kim, Darren Cosker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6056-6066

While there exists a great deal of labeled in-the-wild human data, the same is not true for animals. Manually creating new labels for the full range of animal species would take years of effort from the community. We are also now seeing the emerging potential for computer vision methods in areas like animal conservation, which is an additional motivation for this direction of research. Key to our approach is the ability to easily generate as many labeled training images as we desire across a range of different modalities. To achieve this, we present a new large scale canine motion capture dataset and parametric canine body and texture model. These are used to produce the first large scale, multi-domain, multi-task dataset for canine body analysis comprising of detailed synthetic labels on both real images and fully synthetic images in a range of realistic poses. We also introduce the first pose prior for animals in the form of a variational pose prior for canines which is used to fit the parametric model to images of canines. We demonstrate the effectiveness of our labels for training computer vision models on tasks such as parts-based segmentation and pose estimation and show such models can generalise to other animal species without additional training.

****

Diffusion-Based Generation of Histopathological Whole Slide Images at a Gigapixel Scale

Robert Harb, Thomas Pock, Heimo Müller; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5131-5140

We present a novel diffusion-based approach to generate synthetic histopathological Whole Slide Images (WSIs) at an unprecedented gigapixel scale. Synthetic WSIs have many potential applications: They can augment training datasets to enhance the performance of many computational pathology applications. They allow the creation of synthesized copies of datasets that can be shared without violating privacy regulations. Or they can facilitate learning representations of WSIs without requiring data annotations. Despite this variety of applications, no existing deep-learning-based method generates WSIs at their typically high resolutions. Mainly due to the high computational complexity. Therefore, we propose a novel coarse-to-fine sampling scheme to tackle image generation of high-resolution WSIs. In this scheme, we increase the resolution of an initial low-resolution image to a high-resolution WSI. Particularly, a diffusion model sequentially adds fine details to images and increases their resolution. In our experiments, we train our method with WSIs from the TCGA- BRCA dataset. Additionally to quantitative evaluations, we also performed a user study with pathologists. The study results suggest that our generated WSIs resemble the structure of real WSIs.

****

Bridging the Gap Between Multi-Focus and Multi-Modal: A Focused Integration Framework for Multi-Modal Image Fusion

Xilai Li, Xiaosong Li, Tao Ye, Xiaoqi Cheng, Wuyang Liu, Haishu Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20

24, pp. 1628-1637

Multi-modal image fusion (MMIF) integrates valuable information from different modality images into a fused one. However, the fusion of multiple visible images with different focal regions and infrared images is a unprecedented challenge in real MMIF applications. This is because of the limited depth of the focus of visible optical lenses, which impedes the simultaneous capture of the focal information within the same scene. To address this issue, in this paper, we propose a MMIF framework for joint focused integration and modalities information extraction. Specifically, a semi-sparsity-based smoothing filter is introduced to decompose the images into structure and texture components. Subsequently, a novel multi-scale operator is proposed to fuse the texture components, capable of detecting significant information by considering the pixel focus attributes and relevant data from various modal images. Additionally, to achieve an effective capture of scene luminance and reasonable contrast maintenance, we consider the distribution of energy information in the structural components in terms of multi-directional frequency variance and information entropy. Extensive experiments on existing MMIF datasets, as well as the object detection and depth estimation tasks, consistently demonstrate that the proposed algorithm can surpass the state-of-the-art methods in visual perception and quantitative evaluation.The code is available at https://github.com/ixilai/MFIF-MMIF.

****

Image Labels Are All You Need for Coarse Seagrass Segmentation

Scarlett Raine, Ross Marchant, Brano Kusy, Frederic Maire, Tobias Fischer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5943-5952

Seagrass meadows serve as critical carbon sinks, but estimating the amount of carbon they store requires knowledge of the seagrass species present. Underwater and surface vehicles equipped with machine learning algorithms can help to accurately estimate the composition and extent of seagrass meadows at scale. However, previous approaches for seagrass detection and classification have required supervision from patch-level labels. In this paper, we reframe seagrass classification as a weakly supervised coarse segmentation problem where image-level labels are used during training (25 times fewer labels compared to patch-level labeling) and patch-level outputs are obtained at inference time. To this end, we introduce SeaFeats, an architecture that uses unsupervised contrastive pre-training and feature similarity, and SeaCLIP, a model that showcases the effectiveness of large language models as a supervisory signal in domain-specific applications. We demonstrate that an ensemble of SeaFeats and SeaCLIP leads to highly robust performance. Our method outperforms previous approaches that require patch-level labels on the multi-species 'DeepSeagrass' dataset by 6.8% (absolute) for the class-weighted F1 score, and by 12.1% (absolute) for the seagrass presence/absence F1 score on the 'Global Wetlands' dataset. We also present two case studies for real-world deployment: outlier detection on the Global Wetlands dataset, and application of our method on imagery collected by the FloatyBoat autonomous surface vehicle.

****

Registered and Segmented Deformable Object Reconstruction From a Single View Point Cloud

Pit Henrich, Balázs Gyenes, Paul Maria Scheikl, Gerhard Neumann, Franziska Mathis-Ullrich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3129-3138

In deformable object manipulation, we often want to interact with specific segments of an object that are only defined in non-deformed models of the object. We thus require a system that can recognize and locate these segments in sensor data of deformed real world objects. This is normally done using deformable object registration, which is problem specific and complex to tune. Recent methods utilize neural occupancy functions to improve deformable object registration by registering to an object reconstruction. Going one step further, we propose a system that in addition to reconstruction learns segmentation of the reconstructed object. As the resulting output already contains the information about the segments

, we can skip the registration process. Tested on a variety of deformable objects in simulation and the real world, we demonstrate that our method learns to robustly find these segments. We also introduce a simple sampling algorithm to generate better training data for occupancy learning.

****

Adaptive Manifold for Imbalanced Transductive Few-Shot Learning

Michalis Lazarou, Yannis Avrithis, Tania Stathaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2297-2306

Transductive few-shot learning algorithms have showed substantially superior performance over their inductive counterparts by leveraging the unlabeled queries at inference. However, the vast majority of transductive methods are evaluated on perfectly class-balanced benchmarks. It has been shown that they undergo remarkable drop in performance under a more realistic, imbalanced setting. To this end, we propose a novel algorithm to address imbalanced transductive few-shot learning, named Adaptive Manifold. Our algorithm exploits the underlying manifold of the labeled examples and unlabeled queries by using manifold similarity to predict the class probability distribution of every query. It is parameterized by one centroid per class and a set of manifold parameters that determine the manifold. All parameters are optimized by minimizing a loss function that can be tuned towards class-balanced or imbalanced distributions. The manifold similarity shows substantial improvement over Euclidean distance, especially in the 1-shot setting. Our algorithm outperforms all other state of the art methods in three benchmark datasets, namely miniImageNet, tieredImageNet and CUB, and two different backbones, namely ResNet-18 and WideResNet-28-10. In certain cases, our algorithm outperforms the previous state of the art by as much as 4.2%. The publicly available source code can be found in https://github.com/MichalisLazarou/AM.

****

Restoring Degraded Old Films With Recursive Recurrent Transformer Networks

Shan Lin, Edgar Simo-Serra; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6718-6728

There exists a large number of old films that have not only artistic value but also historical significance. However, due to the degradation of analogue medium over time, old films often suffer from various deteriorations that make it difficult to restore them with existing approaches. In this work, we proposed a novel framework called Recursive Recurrent Transformer Network (RRTN) which is specifically designed for restoring degraded old films. Our approach introduces several key advancements, including a more accurate film noise mask estimation method, the utilization of second-order grid propagation and flow-guided deformable alignment, and the incorporation of a recursive structure to further improve the removal of challenging film noise. Through qualitative and quantitative evaluations, our approach demonstrates superior performance compared to existing approaches, effectively improving the restoration for difficult film noises that cannot be perfectly handled by existing approaches. The code and model are available at https://github.com/mountln/RRTN-old-film-restoration.

****

Re-Evaluating LiDAR Scene Flow

Nathaniel Chodosh, Deva Ramanan, Simon Lucey; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6005-6015

Popular benchmarks for self-supervised LiDAR scene flow (stereoKITTI, and FlyingThings3D) have unrealistic rates of dynamic motion, unrealistic correspondences, and unrealistic sampling patterns. As a result, progress on these benchmarks is misleading and may cause researchers to focus on the wrong problems. We evaluate a suite of top methods on a suite of real-world datasets (Argoverse 2.0, Waymo, and NuScenes) and report several conclusions. First, we find that performance on stereoKITTI is negatively correlated with performance on real-world data. Second, we find that one of this task's key components -- removing the dominant ego-motion -- is better solved by classic ICP than any tested method. Finally, we show that despite the emphasis placed on learning, most performance gains are caused by pre- and post-processing steps: piecewise- rigid refinement and ground removal. We demonstrate this through a baseline method that combines these process ing steps with a learning-free test-time flow optimization. This baseline outperforms every evaluated method

****

Unsupervised 3D Pose Estimation With Non-Rigid Structure-From-Motion Modeling

Haorui Ji, Hui Deng, Yuchao Dai, Hongdong Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3314-3323

Most existing 3D human pose estimation work rely heavily on the powerful memory capability of networks to obtain suitable 2D-3D mappings from the training data. Few works have studied the modeling of human posture deformation in motion. In this paper, we propose a new modeling method for human pose deformations and design an accompanying diffusion-based motion prior. Inspired by the field of non-rigid structure-from-motion, we divide the task of reconstructing 3D human skeletons in motion into the estimation of a 3D reference skeleton, and a frame-by-frame skeleton deformation. A mixed spatial-temporal NRSfMformer is used to simultaneously estimate the 3D reference skeleton and the skeleton deformation of each frame from 2D observations sequence, and then sum them up to obtain the pose of each frame. Subsequently, a loss term based on the diffusion model is used to ensure that the pipeline learns the correct prior motion knowledge. Finally, we have evaluated our proposed method on mainstream datasets and obtained superior results outperforming the state-of-the-art.

****

FAKD: Feature Augmented Knowledge Distillation for Semantic Segmentation

Jianlong Yuan, Minh Hieu Phan, Liyang Liu, Yifan Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 595-605

In this work, we explore data augmentations for knowledge distillation on semantic segmentation. Due the capacity gap, small-sized student networks struggle to discover the discriminative feature space learned by a powerful teacher. Image-level augmentations allow the student to better imitate the teacher by providing extra outputs. However, existing distillation frameworks only augment a limited number of samples, which restricts the learning of a student. Inspired by the recent progress on semantic directions on feature space, this work proposes a feature-level augmented knowledge distillation (FAKD) which infinitely augments features along a semantic direction for optimal knowledge transfer. Furthermore, we introduce novel surrogate loss functions to distill the teacher's knowledge from an infinite number of samples. The surrogate loss is an upper bound of the expected distillation loss over infinite augmented samples. Extensive experiments on four semantic segmentation benchmarks demonstrate that the proposed method boosts the performance of current knowledge distillation methods without any significant overhead. The code will be released at FAKD.

****

TriCoLo: Trimodal Contrastive Loss for Text To Shape Retrieval

Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, Angel X. Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5815-5825

Text-to-shape retrieval is an increasingly relevant problem with the growth of 3D shape data. Recent work on contrastive losses for learning joint embeddings over multimodal data has been successful at tasks such as retrieval and classification. Thus far, work on joint representation learning for 3D shapes and text has focused on improving embeddings through modeling of complex attention between representations, or multi-task learning. We propose a trimodal learning scheme over text, multi-view images and 3D shape voxels, and show that with large batch contrastive learning we achieve good performance on text-to-shape retrieval without complex attention mechanisms or losses. Our experiments serve as a foundation for follow-up work on building trimodal embeddings for text-image-shape.

****

Expanding Expressiveness of Diffusion Models With Limited Data via Self-Distillation Based Fine-Tuning

Jiwan Hur, Jaehyun Choi, Gyojin Han, Dong-Jae Lee, Junmo Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5028-5037

Training diffusion models on limited datasets poses challenges in terms of limited generation capacity and expressiveness, leading to unsatisfactory results in various downstream tasks utilizing pretrained diffusion models, such as domain translation and text-guided image manipulation. In this paper, we propose Self-Distillation for Fine-Tuning diffusion models (SDFT), a methodology to address these challenges by leveraging diverse features from diffusion models pretrained on large source datasets. SDFT distills more general features (shape, colors, etc.) and less domain-specific features (texture, fine details, etc) from the source model, allowing successful knowledge transfer without disturbing the training process on target datasets. The proposed method is not constrained by the specific architecture of the model and thus can be generally adopted to existing frameworks. Experimental results demonstrate that SDFT enhances the expressiveness of the diffusion model with limited datasets, resulting in improved generation capabilities across various downstream tasks.

****

Diff2Lip: Audio Conditioned Diffusion Models for Lip-Synchronization

Soumik Mukhopadhyay, Saksham Suri, Ravi Teja Gadde, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5292-5302

The task of lip synchronization (lip-sync) seeks to match the lips of human faces with different audio. It has various applications in the film industry as well as for creating virtual avatars and for video conferencing. This is a challenging problem as one needs to simultaneously introduce detailed, realistic lip movements while preserving the identity, pose, emotions, and image quality. Many of the previous methods trying to solve this problem suffer from image quality degradation due to a lack of complete contextual information. In this paper, we present Diff2Lip, an audio-conditioned diffusion-based model which is able to do lip synchronization in-the-wild while preserving these qualities. We train our model on Voxceleb2, a video dataset containing in-the-wild talking face videos. Extensive studies show that our method outperforms popular methods like Wav2Lip and PC-AVS in Frechet inception distance (FID) metric and Mean Opinion Scores (MOS) of the users. We show results on both reconstruction (same audio-video inputs) as well as cross (different audio-video inputs) settings on Voxceleb2 and LRWdatasets. Video results are available at https://soumik-kanad.github.io/diff2lip.

****

A*: Atrous Spatial Temporal Action Recognition for Real Time Applications

Myeongjun Kim, Federica Spinola, Philipp Benz, Tae-hoon Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7014-7024

Deep learning has become a popular tool across various fields and is increasingly being integrated into real-world applications such as autonomous driving cars and surveillance cameras. One area of active research is recognizing human actions, including identifying unsafe or abnormal behaviors. Temporal information is crucial for action recognition tasks. Global context, as well as the target person, are also important for judging human behaviors. However, larger networks that can capture all of these features face difficulties operating in real-time. To address these issues, we propose A*: Atrous Spatial Temporal Action Recognition for Real Time Applications. A* includes four modules aimed at improving action detection networks. First, we introduce a Low-Level Feature Aggregation module. Second, we propose the Atrous Spatio-Temporal Pyramid Pooling module. Third, we suggest to fuse all extracted image and video features in an Image-Video Feature Fusion module. Finally, we integrate the Proxy Anchor Loss for action features into the loss function. We evaluate A* on three common action detection benchmarks, and achieve state-of-the-art performance on JHMDB and UCF101-24, while staying competitive on AVA. Furthermore, we demonstrate that A* can achieve real-time inference speeds of 33 FPS, making it suitable for real-world applications.

****

Augment the Pairs: Semantics-Preserving Image-Caption Pair Augmentation for Grounding-Based Vision and Language Models

Jingru Yi, Burak Uzkent, Oana Ignat, Zili Li, Amanmeet Garg, Xiang Yu, Linda Liu

; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5520-5530

Grounding-based vision and language models have been successfully applied to low-level vision tasks, aiming to precisely locate objects referred in captions. The effectiveness of grounding representation learning heavily relies on the scale of the training dataset. Despite being a useful data enrichment strategy, data augmentation has received minimal attention in existing vision and language tasks as augmentation for image-caption pairs is non-trivial. In this study, we propose a robust phrase grounding model trained with text-conditioned and text-unconditioned data augmentations. Specifically, we apply text-conditioned color jittering and horizontal flipping to ensure semantic consistency between images and captions. To guarantee image-caption correspondence in the training samples, we modify the captions according to pre-defined keywords when applying horizontal flipping. Additionally, inspired by recent masked signal reconstruction, we propose to use pixel-level masking as a novel form of data augmentation. While we demonstrate our data augmentation method with MDETR framework, the proposed approach is applicable to common grounding-based vision and language tasks with other frameworks. Finally, we show that larger capacity image encoder such as CLIP can further improve the results. Through extensive experiments on three commonly applied datasets: Flickr30k, referring expressions, and GQA, our method demonstrates advanced performance over the state-of-the-arts with various metrics. Code can be found in https://github.com/amzn/augment-the-pairs-wacv2024.

****

Controllable Image Synthesis of Industrial Data Using Stable Diffusion

Gabriele Valvano, Antonino Agostino, Giovanni De Magistris, Antonino Graziano, Giacomo Veneri; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5354-5363

Training supervised deep neural networks that perform defect detection and segmentation requires large-scale fully-annotated datasets, which can be hard or even impossible to obtain in industrial environments. Generative AI offers opportunities to enlarge small industrial datasets artificially, thus enabling the usage of state-of-the-art supervised approaches in the industry. Unfortunately, also good generative models need a lot of data to train, while industrial datasets are often tiny. Here, we propose a new approach for reusing general-purpose pre-trained generative models on industrial data, ultimately allowing the generation of self-labelled defective images. First, we let the model learn the new concept, entailing the novel data distribution. Then, we force it to learn to condition the generative process, producing industrial images that satisfy well-defined topological characteristics and show defects with a given geometry and location. To highlight the advantage of our approach, we use the synthetic dataset to optimise a crack segmentor for a real industrial use case. When the available data is small, we observe considerable performance increase under several metrics, showing the method's potential in production environments.

****

Understanding Dark Scenes by Contrasting Multi-Modal Observations

Xiaoyu Dong, Naoto Yokoya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 840-850

Understanding dark scenes based on multi-modal image data is challenging, as both the visible and auxiliary modalities provide limited semantic information for the task. Previous methods focus on fusing the two modalities but neglect the correlations among semantic classes when minimizing losses to align pixels with labels, resulting in inaccurate class predictions. To address these issues, we introduce a supervised multi-modal contrastive learning approach to increase the semantic discriminability of the learned multi-modal feature spaces by jointly performing cross-modal and intra-modal contrast under the supervision of the class correlations. The cross-modal contrast encourages same-class embeddings from across the two modalities to be closer and pushes different-class ones apart. The intra-modal contrast forces same-class or different-class embeddings within each modality to be together or apart. We validate our approach on a variety of tasks that cover diverse light conditions and image modalities. Experiments show that our approach can effectively enhance dark scene understanding based on multi-modal images with limited semantics by shaping semantic-discriminative feature spaces. Comparisons with previous methods demonstrate our state-of-the-art performance. Code and pretrained models are available at https://github.com/palmdong/SMMCL.

****

Expanding Hyperspherical Space for Few-Shot Class-Incremental Learning

Yao Deng, Xiang Xiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1967-1976

In today's ever-changing world, the ability of machine learning models to continually learn new data without forgetting previous knowledge is of utmost importance. However, in the scenario of few-shot class-incremental learning (FSCIL), where models have limited access to new instances, this task becomes even more challenging. Current methods use prototypes as a replacement for classifiers, where the cosine similarity of instances to these prototypes is used for prediction. However, we have identified that the embedding space created by using the relu activation function is incomplete and crowded for future classes. To address this issue, we propose the Expanding Hyperspherical Space (EHS) method for FSCIL. In EHS, we utilize an odd-symmetric activation function to ensure the completeness and symmetry of embedding space. Additionally, we specify a region for base classes and reserve space for unseen future classes, which increases the distance between class distributions. Pseudo instances are also used to enable the model to anticipate possible upcoming samples. During inference, we provide rectification to the confidence to prevent bias towards base classes. We conducted experiments on benchmark datasets such as CIFAR100 and miniImageNet, which demonstrate that our proposed method achieves state-of-the-art performance.

****

Differentially Private Video Activity Recognition

Zelun Luo, Yuliang Zou, Yijin Yang, Zane Durante, De-An Huang, Zhiding Yu, Chaowei Xiao, Li Fei-Fei, Animashree Anandkumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6657-6667

In recent years, differential privacy has seen significant advancements in image classification; however, its application to video activity recognition remains under-explored. This paper addresses the challenges of applying differential privacy to video activity recognition, which primarily stem from: (1) a discrepancy between the desired privacy level for entire videos and the nature of input data processed by contemporary video architectures, which are typically short, segmented clips; and (2) the complexity and sheer size of video datasets relative to those in image classification, which render traditional differential privacy methods inadequate. To tackle these issues, we propose Multi-Clip DP-SGD, a novel framework for enforcing video-level differential privacy through clip-based classification models. This method samples multiple clips from each video, averages their gradients, and applies gradient clipping in DP-SGD without incurring additional privacy loss. Moreover, we incorporate a parameter-efficient transfer learning strategy to make the model scalable for large-scale video datasets. Through extensive evaluations on the UCF-101 and HMDB-51 datasets, our approach exhibits impressive performance, achieving 81% accuracy with a privacy budget of epsilon=5 on UCF-101, marking a 76% improvement compared to a direct application of DP-SGD. Furthermore, we demonstrate that our transfer learning strategy is versatile and can enhance differentially private image classification across an array of datasets including CheXpert, ImageNet, CIFAR-10, and CIFAR-100.

****

Towards a Dynamic Vision Sensor-Based Insect Camera Trap

Eike Gebauer, Sebastian Thiele, Pierre Ouvrard, Adrien Sicard, Benjamin Risse; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7157-7166

This paper introduces a visual real-time insect monitoring approach capable of detecting and tracking tiny and fast-moving objects in cluttered wildlife conditions using an RGB-DVS stereo-camera system. By building on the intrinsic benefits of event vision data acquisition, we demonstrate that insect presence can be de tected at an extremely high temporal rate (on average more than 40 times real-time) while surpassing the spatial and spectral sensitivity of conventional colour-based sensing. Our DVS-based detection and tracking algorithm extracts insect locations over time, and we evaluated our system based on 81104 manually annotated stereo-frames with 34453 insect appearances featuring highly varying scenes and imaging conditions (including clutter, wind-induced motion, etc.). Comparing our algorithm to two state-of-the-art deep learning algorithms reveals superior results in both detection performance and computational speed. Using the DVS as a trigger for the temporally synchronised RGB camera, we are able to correctly identify 73% of images with and without insects which can be increased to 76% with parameters optimised for different scenes. Overall, our study suggests that DVS-based sensing can be used for visual insect monitoring by enabling reliable real-time insect detection in wildlife conditions while significantly reducing the necessity for data storage, manual labour and energy.

****

FLORA: Fine-Grained Low-Rank Architecture Search for Vision Transformer

Chi-Chih Chang, Yuan-Yao Sung, Shixing Yu, Ning-Chi Huang, Diana Marculescu, Kai-Chiang Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2482-2491

Vision Transformers (ViT) have recently demonstrated success across a myriad of computer vision tasks. However, their elevated computational demands pose significant challenges for real-world deployment. While low-rank approximation stands out as a renowned method to reduce computational loads, efficiently automating the target rank selection in ViT remains a challenge. Drawing from the notable similarity and alignment between the processes of rank selection and One-Shot NAS, we introduce FLORA, an end-to-end automatic framework based on NAS. To overcome the design challenge of supernet posed by vast search space, FLORA employs a low-rank aware candidate filtering strategy. This method adeptly identifies and eliminates underperforming candidates, effectively alleviating potential undertraining and interference among subnetworks. To further enhance the quality of low-rank supernets, we design a low-rank specific training paradigm. First, we propose weight inheritance to construct supernet and enable gradient sharing among low-rank modules. Secondly, we adopt low-rank aware sampling to strategically allocate training resources, taking into account inherited information from pre-trained models. Empirical results underscore FLORA's efficacy. With our method, a more fine-grained rank configuration can be generated automatically and yield up to 33% extra FLOPs reduction compared to a simple uniform configuration. More specific, FLORA-DeiT-B/FLORA-Swin-B can save up to 55%/42% FLOPs almost without performance degradtion. Importantly, FLORA boasts both versatility and orthogonality, offering an extra 21%-26% FLOPs reduction when integrated with leading compression techniques or compact hybrid structures. Our code is publicly available at https://github.com/shadowpa0327/FLORA.

****

Latent-Guided Exemplar-Based Image Re-Colorization

Wenjie Yang, Ning Xu, Yifei Fan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4250-4259

Exemplar-based re-colorization transfers colors from a reference to a colored or grayscale source image, accounting for the semantic correspondences between the two. Existing grayscale colorization methods usually predict only the chromatic aberration while maintaining the source's luminance. Consequently, the result's color may diverge from the reference due to such luminance difference. On the other hand, global photorealistic stylization without segmentation cannot handle scenarios where different parts of the scene need different colors. To overcome this issue, we propose a novel and effective method for re-colorization: 1) We first exploit the spatial-adaptive latent space of SpaceEdit in the context of the re-colorization task and achieve re-colorization via latent maps prediction through a proposed network. 2) We then delve into SpaceEdit's self-reconstruct latent codes and maps to better characterize the global style and local color property, based on which we construct a novel loss to supervise re-colorization. Qualitative and quantitative results show that our method outperforms previous works by generating superior outputs with more consistent colors and global styles based on references.

****

Data Augmentation for Object Detection via Controllable Diffusion Models

Haoyang Fang, Boran Han, Shuai Zhang, Su Zhou, Cuixiong Hu, Wen-Ming Ye; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1257-1266

Data augmentation is vital for object detection tasks that require expensive bounding box annotations. Recent successes in diffusion models have inspired the use of diffusion-based synthetic images for data augmentation. However, existing works have primarily focused on image classification, and their applicability to boost object detection's performance remains unclear. To address this gap, we propose a data augmentation pipeline based on controllable diffusion models and CLIP. Our approach involves generating appropriate visual priors to control the generation of synthetic data and implementing post-filtering techniques using category-calibrated CLIP scores. The evaluation of our approach is conducted under few-shot settings in MSCOCO, full PASCAL VOC dataset, and selected downstream datasets. We observe the performance increase using our augmentation pipeline. Specifically, the mAP improvement is +18.0%/+15.6%/+15.9% for COCO 5/10/30-shot, +2.9% on full PASCAL VOC dataset, and +12.4% on average for selected downstream datasets.

****

Self-Supervised Learning With Masked Autoencoders for Teeth Segmentation From Intra-Oral 3D Scans

Amani Almalki, Longin Jan Latecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7820-7830

In modern dentistry, teeth localization, segmentation, and labeling from intra-oral 3D scans are crucial for improving dental diagnostics, treatment planning, and population-based studies on oral health. However, creating automated algorithms for teeth analysis is a challenging task due to the limited availability of accessible data for training, particularly from the point of view of deep learning. This study extends the self-supervised learning framework of the mesh masked autoencoder (MeshMAE) transformer. While the MeshMAE loss measures the quality of reconstructed masked mesh triangles, the loss of the proposed DentalMAE evaluates the predicted deep embeddings of masked mesh triangles. This yields a better generalization ability on a very limited number of 3D dental scans, as documented by our results on teeth segmentation of intra-oral scans. Our results show that masking-based unsupervised learning methods may, for the first time, provide convincing transfer learning improvements on 3D intra-oral scans, increasing the overall accuracy over both MeshMAE and prior self-supervised pre-training.

****

Small Objects Matters in Weakly-Supervised Semantic Segmentation

Cheolhyun Mun, Sanghuk Lee, Youngjung Uh, Junsuk Choe, Hyeran Byun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 414-423

Weakly-supervised semantic segmentation (WSSS) performs pixel-wise classification given only image-level labels for training. Despite the difficulty of this task, the research community has achieved promising results over the last five years. Still, current WSSS literature misses the detailed sense of how well the methods perform on different sizes of objects. Thus we propose a novel evaluation metric to provide a comprehensive assessment across different object sizes and collect a size-balanced evaluation set to complement PASCAL VOC. With these two gadgets, we reveal that the existing WSSS methods struggle in capturing small objects. Furthermore, we propose a size-balanced cross-entropy loss coupled with a proper training strategy. It generally improves existing WSSS methods as validated upon ten baselines on three different datasets.

****

MaskConver: Revisiting Pure Convolution Model for Panoptic Segmentation

Abdullah Rashwan, Jiageng Zhang, Ali Taalimi, Fan Yang, Xingyi Zhou, Chaochao Yan, Liang-Chieh Chen, Yeqing Li; Proceedings of the IEEE/CVF Winter Conference on

Applications of Computer Vision (WACV), 2024, pp. 851-861

In recent years, transformer-based models have dominated panoptic segmentation, thanks to their strong modeling capabilities and their unified representation for both semantic and instance classes as global binary masks. In this paper, we revisit pure convolution model and propose a novel panoptic architecture named MaskConver. MaskConver proposes to fully unify things and stuff representation by predicting their centers. To that extent, it creates a lightweight class embedding module that can break the ties when multiple centers co-exist in the same location. Furthermore, our study shows that the decoder design is critical in ensuring that the model has sufficient context for accurate detection and segmentation. We introduce a powerful ConvNeXt-UNet decoder that closes the performance gap between convolution- and transformer-based models. With ResNet50 backbone, our MaskConver achieves 53.6% PQ on the COCO panoptic val set, out-performing the modern convolution-based model, Panoptic FCN, by 9.3% as well as transformer-based models such as Mask2Former (+1.7% PQ) and kMaX-DeepLab (+0.6% PQ). Additionally, MaskConver with a MobileNet backbone reaches 37.2% PQ, improving over Panoptic-DeepLab by +6.4% under the same FLOPs/latency constraints. A further optimized version of MaskConver achieves 29.7% PQ, while running in real-time on mobile devices. The code and model weights will be publicly available.

****

From Chaos to Calibration: A Geometric Mutual Information Approach To Target-Free Camera LiDAR Extrinsic Calibration

Jack Borer, Jeremy Tschirner, Florian Ölsner, Stefan Milz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8409-8418

Sensor fusion is vital for the safe and robust operation of autonomous vehicles. Accurate extrinsic sensor to sensor calibration is necessary to accurately fuse multiple sensor's data in a common spatial reference frame. In this paper, we propose a target free extrinsic calibration algorithm that requires no ground truth training data, artificially constrained motion trajectories, hand engineered features or offline optimization and that is accurate, precise and extremely robust to initialization error. Most current research on online camera-LiDAR extrinsic calibration requires ground truth training data which is impossible to capture at scale. We revisit analytical mutual information based methods first proposed in 2012 and demonstrate that geometric features provide a robust information metric for camera-LiDAR extrinsic calibration. We demonstrate our proposed improvement using the KITTI and KITTI-360 fisheye data set.

****

PHG-Net: Persistent Homology Guided Medical Image Classification

Yaopeng Peng, Hongxiao Wang, Milan Sonka, Danny Z. Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7583-7592

Modern deep neural networks have achieved great successes in medical image analysis. However, the features captured by convolutional neural networks (CNNs) or Transformers tend to be optimized for pixel intensities and neglect key anatomical structures such as connected components and loops. In this paper, we propose a persistent homology guided approach (PHG-Net) that explores topological features of objects for medical image classification. For an input image, we first compute its cubical persistence diagram and extract topological features into a vector representation using a small neural network (called the PH module). The extracted topological features are then incorporated into the feature map generated by CNN or Transformer for feature fusion. The PH module is lightweight and capable of integrating topological features into any CNN or Transformer architectures in an end-to-end fashion. We evaluate our PHG-Net on three public datasets and demonstrate its considerable improvements on the target classification tasks over state-of-the-art methods.

****

Masking Improves Contrastive Self-Supervised Learning for ConvNets, and Saliency Tells You Where

Zhi-Yi Chin, Chieh-Ming Jiang, Ching-Chun Huang, Pin-Yu Chen, Wei-Chen Chiu; Pro ceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2761-2770

While image data starts to enjoy the simple-but-effective self-supervised learning scheme built upon masking and self-reconstruction objective thanks to the introduction of tokenization procedure and vision transformer backbone, convolutional neural networks as another important and widely-adopted architecture for image data, though having contrastive-learning techniques to drive the self-supervised learning, still face the difficulty of leveraging such straightforward and general masking operation to benefit their learning process significantly. In this work, we aim to alleviate the burden of including masking operation into the contrastive-learning framework for convolutional neural networks as an extra augmentation method. In addition to the additive but unwanted edges (between masked and unmasked regions) as well as other adverse effects caused by the masking operations for ConvNets, which have been discussed by prior works, we particularly identify the potential problem where for one view in a contrastive sample-pair the randomly-sampled masking regions could be overly concentrated on important/salient objects thus resulting in misleading contrastiveness to the other view. To this end, we propose to explicitly take the saliency constraint into consideration in which the masked regions are more evenly distributed among the foreground and background for realizing the masking-based augmentation. Moreover, we introduce hard negative samples by masking larger regions of salient patches in an input image. Extensive experiments conducted on various datasets, contrastive learning mechanisms, and downstream tasks well verify the efficacy as well as the superior performance of our proposed method with respect to several state-of-the-art baselines.

****

Cheating Depth: Enhancing 3D Surface Anomaly Detection via Depth Simulation

Vitjan Zavrtanik, Matej Kristan, Danijel Sko■aj; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2164-2172

RGB-based surface anomaly detection methods have advanced significantly. However, certain surface anomalies remain practically invisible in RGB alone, necessitating the incorporation of 3D information. Existing approaches that employ point-cloud backbones suffer from suboptimal representations and reduced applicability due to slow processing. Re-training RGB backbones, designed for faster dense input processing, on industrial depth datasets is hindered by the limited availability of sufficiently large datasets. We make several contributions to address these challenges. (i) We propose a novel Depth-Aware Discrete Autoencoder (DADA) architecture, that enables learning a general discrete latent space that jointly models RGB and 3D data for 3D surface anomaly detection. (ii) We tackle the lack of diverse industrial depth datasets by introducing a simulation process for learning informative depth features in the depth encoder. (iii) We propose a new surface anomaly detection method 3DSR, which outperforms all existing state-of-the-art on the challenging MVTec3D anomaly detection benchmark, both in terms of accuracy and processing speed. The experimental results validate the effectiveness and efficiency of our approach, highlighting the potential of utilizing depth information for improved surface anomaly detection.

****

CLID: Controlled-Length Image Descriptions With Limited Data

Elad Hirsch, Ayellet Tal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5531-5541

Controllable image captioning models generate human-like image descriptions, enabling some kind of control over the generated captions. This paper focuses on controlling the caption length, i.e. a short and concise description or a long and detailed one. Since existing image captioning datasets contain mostly short captions, generating long captions is challenging. To address the shortage of long training examples, we propose to enrich the dataset with varying-length self-generated captions. These, however, might be of varying quality and are thus unsuitable for conventional training. We introduce a novel training strategy that selects the data points to be used at different times during the training. Our method dramatically improves the length-control abilities, while exhibiting SoTA perf ormance in terms of caption quality. Our approach is general and is shown to be applicable also to paragraph generation.

****

Steering Prototypes With Prompt-Tuning for Rehearsal-Free Continual Learning

Zhuowei Li, Long Zhao, Zizhao Zhang, Han Zhang, Di Liu, Ting Liu, Dimitris N. Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2523-2533

In the context of continual learning, prototypes--as representative class embeddings--offer advantages in memory conservation and the mitigation of catastrophic forgetting. However, challenges related to semantic drift and prototype interference persist. In this study, we introduce the Contrastive Prototypical Prompt (CPP) approach. Through task-specific prompt-tuning, underpinned by a contrastive learning objective, we effectively address both aforementioned challenges. Our evaluations on four challenging class-incremental benchmarks reveal that CPP achieves a significant 4% to 6% improvement over state-of-the-art methods. Importantly, CPP operates without a rehearsal buffer and narrows the performance divergence between continual and offline joint learning, suggesting an innovative scheme for Transformer-based continual learning systems.

****

Modality-Aware Representation Learning for Zero-Shot Sketch-Based Image Retrieval

Eunyi Lyou, Doyeon Lee, Jooeun Kim, Joonseok Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5646-5655

Zero-shot learning offers an efficient solution for a machine learning model to treat unseen categories, avoiding exhaustive data collection. Zero-shot Sketch-based Image Retrieval (ZS-SBIR) simulates real-world scenarios where it is hard and costly to collect paired sketch-photo samples. We propose a novel framework that indirectly aligns sketches and photos by contrasting them through texts, removing the necessity of access to sketch-photo pairs. With an explicit modality encoding learned from data, our approach disentangles modality-agnostic semantics from modality-specific information, bridging the modality gap and enabling effective cross-modal content retrieval within a joint latent space. From comprehensive experiments, we verify the efficacy of the proposed model on ZS-SBIR, and it can be also applied to generalized and fine-grained settings.

****

Concurrent Band Selection and Traversability Estimation From Long-Wave Hyperspectral Imagery in Off-Road Settings

Florence Yellin, Scott McCloskey, Cole Hill, Eric Smith, Brian Clipp; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7483-7492

Autonomous navigation has become increasingly popular in recent years; However, most existing methods focus on on-road navigation and utilize active sensors, such as LiDAR. This paper instead focuses on autonomous off-road navigation using traversability estimation from passive sensors, specifically long-wave (LW) hyperspectral imagery (HSI). We present a method for selecting a subset of hyperspectral bands that are most useful for traversability estimation by designing a band selection module that designs a minimal sensor that measures sparsely-sampled spectral bands while jointly training a semantic segmentation network for traversability estimation. The effectiveness of our method is demonstrated using our dataset of LW HSI from diverse off-road scenes including forest, desert, snow, ponds, and open fields. Our dataset includes imagery collected both during the daytime and nighttime during various weather conditions, including challenging scenes with a wide range of obstacles. Using our method, we learn a small subset (2%) of all the HSI bands that can achieve competitive or better traversability estimation accuracy to that achieved when utilizing all hyperspectral bands. Using only 5 bands, our method is able to achieve a mean class accuracy that is only 1.3% less than that achieved using full 256-band HSI and only 0.1% less than that achieved using 250-band HSI, demonstrating the success of our method.

****

Token Fusion: Bridging the Gap Between Token Pruning and Token Merging

Minchul Kim, Shangqian Gao, Yen-Chang Hsu, Yilin Shen, Hongxia Jin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1383-1392

Vision Transformers (ViTs) have emerged as powerful backbones in computer vision, outperforming many traditional CNNs. However, their computational overhead, largely attributed to the self-attention mechanism, makes deployment on resource-constrained edge devices challenging. Multiple solutions rely on token pruning or token merging. In this paper, we introduce "Token Fusion" (ToFu), a method that amalgamates the benefits of both token pruning and token merging. Token pruning proves advantageous when the model exhibits sensitivity to input interpolations, while token merging is effective when the model manifests close to linear responses to inputs. We combine this to propose a new scheme called Token Fusion. Moreover, we tackle the limitations of average merging, which doesn't preserve the intrinsic feature norm, resulting in distributional shifts. To mitigate this, we introduce MLERP merging, a variant of the SLERP technique, tailored to merge multiple tokens while maintaining the norm distribution. ToFu is versatile, applicable to ViTs with or without additional training. Our empirical evaluations indicate that ToFu establishes new benchmarks in both classification and image generation tasks concerning computational efficiency and model accuracy.

****

Global Occlusion-Aware Transformer for Robust Stereo Matching

Zihua Liu, Yizhou Li, Masatoshi Okutomi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3535-3544

Despite the remarkable progress facilitated by learning-based stereo-matching algorithms, the performance in the ill-conditioned regions, such as the occluded regions, remains a bottleneck. Due to the limited receptive field, existing CNN-based methods struggle to handle these ill-conditioned regions effectively. To address this issue, this paper introduces a novel attention-based stereo-matching network called Global Occlusion-Aware Transformer (GOAT) to exploit long-range dependency and occlusion-awareness global context for disparity estimation. In the GOAT architecture, a parallel disparity and occlusion estimation module PDO is proposed to estimate the initial disparity map and the occlusion mask using a parallel attention mechanism. To further enhance the disparity estimates in the occluded regions, an occlusion-aware global aggregation module (OGA) is proposed. This module aims to refine the disparity in the occluded regions by leveraging restricted global correlation within the focus scope of the occluded areas. Extensive experiments were conducted on several public benchmark datasets including SceneFlow, KITTI 2015, and Middlebury. The results show that the proposed GOAT demonstrates outstanding performance among all benchmarks, particularly in the occluded regions.

****

SGRec3D: Self-Supervised 3D Scene Graph Learning via Object-Level Scene Reconstruction

Sebastian Koch, Pedro Hermosilla, Narunas Vaskevicius, Mirco Colosi, Timo Ropinski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3404-3414

In the field of 3D scene understanding, 3D scene graphs have emerged as a new scene representation that combines geometric and semantic information about objects and their relationships. However, learning semantic 3D scene graphs in a fully supervised manner is inherently difficult as it requires not only object-level annotations but also relationship labels. While pre-training approaches have helped to boost the performance of many methods in various fields, pre-training for 3D scene graph prediction has received little attention. Furthermore, we find in this paper that classical contrastive point cloud-based pre-training approaches are ineffective for 3D scene graph learning. To this end, we present SGRec3D, a novel self-supervised pre-training method for 3D scene graph prediction. We propose to reconstruct the 3D input scene from a graph bottleneck as a pretext task. Pre-training SGRec3D does not require object relationship labels, making it possible to exploit large-scale 3D scene understanding datasets, which were off-limits for 3D scene graph learning before. Our experiments demonstrate that in co ntrast to recent point cloud-based pre-training approaches, our proposed pre-training improves the 3D scene graph prediction considerably, which results in SOTA performance, outperforming other 3D scene graph models by +10% on object prediction and +4% on relationship prediction. Additionally, we show that only using a small subset of 10% labeled data during fine-tuning is sufficient to outperform the same model without pre-training.

****

Estimating Fog Parameters From an Image Sequence Using Non-Linear Optimisation

Yining Ding, Andrew M. Wallace, Sen Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1578-1586

Given a sequence of images taken in foggy weather, we seek to estimate the atmospheric light and the scattering coefficient. These are key parameters to characterise the nature of the fog, to reconstruct a clear image (defogging), and to infer scene depth. Existing methods adopt a sequential estimation strategy which is prone to error propagation. In sharp contrast, we take a more systematic approach and jointly estimate these parameters by solving a unified non-linear optimisation problem. Experimental results show that the proposed method is superior to existing ones in terms of both estimation accuracy and precision. Our method further demonstrates how image defogging and depth estimation can be linked to a visual localisation system, contributing to more comprehensive and robust perception in fog.

****

Fast and Interpretable Face Identification for Out-of-Distribution Data Using Vision Transformers

Hai Phan, Cindy X. Le, Vu Le, Yihui He, Anh “Totti” Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6301-6311

Most face identification approaches employ a Siamese neural network to compare two images at the image embedding level. Yet, this technique can be subject to occlusion (e.g., faces with masks or sunglasses) and out-of-distribution data. DeepFace-EMD (Phan et al. 2022) reaches state-of-the-art accuracy on out-of-distribution data by first comparing two images at the image level, and then at the patch level. Yet, its later patch-wise re-ranking stage admits a large O(n^3 log n) time complexity (for n patches in an image) due to the optimal transport optimization. In this paper, we propose a novel, 2-image Vision Transformers (ViTs) that compares two images at the patch level using cross-attention. After training on 2M pairs of images on CASIA Webface (Yi et al. 2014), our model performs at a comparable accuracy as DeepFace-EMD on out-of-distribution data, yet at an inference speed more than twice as fast as DeepFace-EMD (Phan et al. 2022). In addition, via a human study, our model shows promising explainability through the visualization of cross-attention. We believe our work can inspire more explorations in using ViTs for face identification.

****

Investigating the Role of Attribute Context in Vision-Language Models for Object Recognition and Detection

Kyle Buettner, Adriana Kovashka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5474-5484

Vision-language alignment learned from image-caption pairs has been shown to benefit tasks like object recognition and detection. Methods are mostly evaluated in terms of how well object class names are learned, but captions also contain rich attribute context that should be considered when learning object alignment. It is unclear how methods use this context in learning, as well as whether models succeed when tasks require attribute and object understanding. To address this gap, we conduct extensive analysis of the role of attributes in vision-language models. We specifically measure model sensitivity to the presence and meaning of attribute context, gauging influence on object embeddings through unsupervised phrase grounding and classification via description methods. We further evaluate the utility of attribute context in training for open-vocabulary object detection, fine-grained text-region retrieval, and attribution tasks. Our results show that attribute context can be wasted when learning alignment for detection, attr ibute meaning is not adequately considered in embeddings, and describing classes by only their attributes is ineffective. A viable strategy that we find to increase benefits from attributes is contrastive training with adjective-based negative captions.

****

Membership Inference Attack Using Self Influence Functions

Gilad Cohen, Raja Giryes; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4892-4901

Member inference (MI) attacks aim to determine if a specific data sample was used to train a machine learning model. Thus, MI is a major privacy threat to models trained on private sensitive data, such as medical records. In MI attacks one may consider the black-box settings, where the model's parameters and activations are hidden from the adversary, or the white-box case where they are available to the attacker. In this work, we focus on the latter and present a novel MI attack for it that employs influence functions, or more specifically the samples' self-influence scores, to perform MI prediction. The proposed method is evaluated on CIFAR-10, CIFAR-100, and Tiny ImageNet datasets using various architectures such as AlexNet, ResNet, and DenseNet. Our new attack method achieves new state-of-the-art (SOTA) results for MI even with limited adversarial knowledge, and is effective against MI defense methods such as data augmentation and differential privacy. Our code is available at https: //github.com/giladcohen/sif_mi_attack.

****

Mixing Gradients in Neural Networks as a Strategy To Enhance Privacy in Federated Learning

Shaltiel Eloul, Fran Silavong, Sanket Kamthe, Antonios Georgiadis, Sean J. Moran; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3956-3965

Federated learning reduces the risk of information leakage, but remains vulnerable to attack. We show that well-mixed gradients provide numerical resistance to gradient inversion in neural networks. For example, we can enhance mixing gradients in a batch by choosing an appropriate loss function and drawing identical labels, and we support this with an approximate solution of batch inversion for linear layers. These simple architecture choices show no degradation to classification performance as opposed to noise perturbation defense. To accurately assess data recovery, we propose to use a variation distance metric for information leakage in images, derived from total variation. In contrast to Mean Squared Error or Structural Similarity Index metrics, it provides a continuous metric for information recovery. Finally, our empirical results of information recovery from various inversion attacks and training performance supports our defense strategies. These simple architecture choices found to be also useful for practical size of convolutional neural networks but depends on their size. We hope this work will trigger further defense studies using gradient mixing, towards achieving a trustful federation policy.

****

Learning to Read Analog Gauges from Synthetic Data

Juan Leon-Alcazar, Yazeed Alnumay, Cheng Zheng, Hassane Trigui, Sahejad Patel, Bernard Ghanem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8616-8625

Manually reading and logging gauge data is time-inefficient, and the effort increases according to the number of gauges available. We present a pipeline that automates the reading of analog gauges. We propose a two-stage CNN pipeline that identifies the key structural components of an analog gauge and outputs an angular reading. To facilitate the training of our approach, a synthetic dataset is generated thus obtaining a set of realistic analog gauges with their corresponding annotation. To validate our proposal, an additional real-world dataset was collected with 4.813 manually curated images. When compared against state-of-the-art methodologies, our method shows a significant improvement of 4.55 in the average error, which is a 52% relative improvement. The resources for this project will be made available at: https://github.com/fuankarion/automatic-gauge-reading.

****

Learning Saliency From Fixations

Yasser Abdelaziz Dahou Djilali, Kevin McGuinness, Noel O’Connor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 383-393

We present a novel approach for saliency prediction in images, leveraging parallel decoding in transformers to learn saliency solely from fixation maps. Models typically rely on continuous saliency maps, to overcome the difficulty of optimizing for the discrete fixation map. We attempt to replicate the experimental setup that generates saliency datasets. Our approach treats saliency prediction as a direct set prediction problem, via a global loss that enforces unique fixations prediction through bipartite matching and a transformer encoder-decoder architecture. By utilizing a fixed set of learned fixation queries, the cross-attention reasons over the image features to directly output the fixation points, distinguishing it from other modern saliency predictors. Our approach, named Saliency TRansformer (SalTR) achieves remarkable results on the Salicon benchmark.

****

PECoP: Parameter Efficient Continual Pretraining for Action Quality Assessment

Amirhossein Dadashzadeh, Shuchao Duan, Alan Whone, Majid Mirmehdi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 42-52

The limited availability of labelled data in Action Quality Assessment (AQA), has forced previous works to fine-tune their models pretrained on large-scale domain-general datasets. This common approach results in weak generalisation, particularly when there is a significant domain shift. We propose a novel, parameter efficient, continual pretraining framework, PECoP, to reduce such domain shift via an additional pretraining stage. In PECoP, we introduce 3D-Adapters, inserted into the pretrained model, to learn spatiotemporal, in-domain information via self-supervised learning where only the adapter modules' parameters are updated. We demonstrate PECoP's ability to enhance the performance of recent state-of-the-art methods (MUSDL, CoRe, and TSA) applied to AQA, leading to considerable improvements on benchmark datasets, JIGSAWS (| 6.0%), MTL-AQA (| 0.99%), and FineDiving (| 2.54%). We also present a new Parkinson's Disease dataset, PD4T, of real patients performing four various actions, where we surpass (| 3.56%) the state-of-the-art in comparison. Our code, pretrained models, and the PD4T dataset are available at https://github.com/Plrbear/PECoP.

****

Face Identity-Aware Disentanglement in StyleGAN

Adrian Suwa■a, Bartosz Wójcik, Magdalena Proszewska, Jacek Tabor, Przemys■aw Spurek, Marek ■mieja; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5222-5231

Conditional GANs are frequently used for manipulating the attributes of face images, such as expression, hairstyle, pose, or age. Even though the state-of-the-art models successfully modify the requested attributes, they simultaneously modify other important characteristics of the image, such as a person's identity. In this paper, we focus on solving this problem by introducing PluGeN4Faces, a plugin to StyleGAN, which explicitly disentangles face attributes from a person's identity. Our key idea is to perform training on images retrieved from movie frames, where a given person appears in various poses and with different attributes. By applying a type of contrastive loss, we encourage the model to group images of the same person in similar regions of latent space. Our experiments demonstrate that the modifications of face attributes performed by PluGeN4Faces are significantly less invasive on the remaining characteristics of the image than in the existing state-of-the-art models.

****

A Robust Diffusion Modeling Framework for Radar Camera 3D Object Detection

Zizhang Wu, Yunzhe Wu, Xiaoquan Wang, Yuanzhu Gan, Jian Pu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3282-3292

Radar-camera 3D object detection aims at interacting radar signals with camera images for identifying objects of interest and localizing their corresponding 3D bounding boxes. To overcome the severe sparsity and ambiguity of radar signals, we propose a robust framework based on probabilistic denoising diffusion modeling. We design our framework to be easily implementable on different multi-view 3D detectors without the requirement of using LiDAR point clouds during either the training or inference. In specific, we first design our framework with a denoised radar-camera encoder via developing a lightweight denoising diffusion model with semantic embedding. Secondly, we develop the query denoising training into 3D space via introducing the reconstruction training at depth measurement for the transformer detection decoder. Our framework achieves new state-of-the-art performance on the nuScenes 3D detection benchmark but with few computational cost increases compared to the baseline detectors.

****

InfraParis: A Multi-Modal and Multi-Task Autonomous Driving Dataset

Gianni Franchi, Marwane Hariat, Xuanlong Yu, Nacim Belkhir, Antoine Manzanera, David Filliat; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2973-2983

Current deep neural networks (DNNs) for autonomous driving computer vision are typically trained on specific datasets that only involve a single type of data and urban scenes. Consequently, these models struggle to handle new objects, noise, nighttime conditions, and diverse scenarios, which is essential for safety-critical applications. Despite ongoing efforts to enhance the resilience of computer vision DNNs, progress has been sluggish, partly due to the absence of benchmarks featuring multiple modalities. We introduce a novel and versatile dataset named InfraParis that supports multiple tasks across three modalities: RGB, depth, and infrared. We assess various state-of-the-art baseline techniques, encompassing models for the tasks of semantic segmentation, object detection, and depth estimation.

****

LAVSS: Location-Guided Audio-Visual Spatial Audio Separation

Yuxin Ye, Wenming Yang, Yapeng Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5508-5519

Existing machine learning research has achieved promising results in monaural audio-visual separation (MAVS). However, most MAVS methods purely consider what the sound source is, not where it is located. This can be a problem in VR/AR scenarios, where listeners need to be able to distinguish between similar audio sources located in different directions. To address this limitation, we have generalized MAVS to spatial audio separation and proposed LAVSS: a location-guided audio-visual spatial audio separator. LAVSS is inspired by the correlation between spatial audio and visual location. We introduce the phase difference carried by binaural audio as spatial cues, and we utilize positional representations of sounding objects as additional modality guidance. We also leverage multi-level cross-modal attention to perform visual-positional collaboration with audio features. In addition, we adopt a pre-trained monaural separator to transfer knowledge from rich mono sounds to boost spatial audio separation. This exploits the correlation between monaural and binaural channels. Experiments on the FAIR-Play dataset demonstrate the superiority of the proposed LAVSS over existing benchmarks of audio-visual separation. Our project page: https://yyx666660.github.io/LAVSS/.

****

PIDiffu: Pixel-Aligned Diffusion Model for High-Fidelity Clothed Human Reconstruction

Jungeun Lee, Sanghun Kim, Hansol Lee, Tserendorj Adiya, Hwasup Lim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5172-5181

This paper presents the Pixel-aligned Diffusion Model (PIDiffu), a new framework for reconstructing high-fidelity clothed 3D human models from a single image. While existing PIFu variants have made significant advances using more complicated 2D and 3D feature extractions, these methods still suffer from floating artifacts and body part duplication due to their reliance on point-wise occupancy field estimations. PIDiffu employs a diffusion-based strategy for line-wise estimation along the ray direction, conditioned by pixel-aligned features with a guided attention. This approach improves the local details and structural accuracy of the reconstructed body shape and is robust to unfamiliar and complex image features. Moreover, PIDiffu can be easily integrated with existing PIFu-based methods to leverage their advantages. The paper demonstrates that PIDiffu outperforms state-of-the-art methods that do not rely on parametric 3D body models. Especially, our method is superior in handling 'in-the-wild' images, such as those with complex patterned clothes unseen in the training data.

****

Kaizen: Practical Self-Supervised Continual Learning With Continual Fine-Tuning

Chi Ian Tang, Lorena Qendro, Dimitris Spathis, Fahim Kawsar, Cecilia Mascolo, Akhil Mathur; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2841-2850

Self-supervised learning (SSL) has shown remarkable performance in computer vision tasks when trained offline. However, in a Continual Learning (CL) scenario where new data is introduced progressively, models still suffer from catastrophic forgetting. Retraining a model from scratch to adapt to newly generated data is time-consuming and inefficient. Previous approaches suggested re-purposing self-supervised objectives with knowledge distillation to mitigate forgetting across tasks, assuming that labels from all tasks are available during fine-tuning. In this paper, we generalize self-supervised continual learning in a practical setting where available labels can be leveraged in any step of the SSL process. With an increasing number of continual tasks, this offers more flexibility in the pre-training and fine-tuning phases. With Kaizen, we introduce a training architecture that is able to mitigate catastrophic forgetting for both the feature extractor and classifier with a carefully designed loss function. By using a set of comprehensive evaluation metrics reflecting different aspects of continual learning, we demonstrated that Kaizen significantly outperforms previous SSL models in competitive vision benchmarks, with up to 16.5% accuracy improvement on split CIFAR-100. Kaizen is able to balance the trade-off between knowledge retention and learning from new data with an end-to-end model, paving the way for practical deployment of continual learning systems.

****

SBCFormer: Lightweight Network Capable of Full-Size ImageNet Classification at 1 FPS on Single Board Computers

Xiangyong Lu, Masanori Suganuma, Takayuki Okatani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1123-1133

Computer vision has become increasingly prevalent in solving real-world problems across diverse domains, including smart agriculture, fishery, and livestock management. These applications may not require processing many image frames per second, leading practitioners to use single board computers (SBCs). Although many lightweight networks have been developed for "mobile/edge" devices, they primarily target smartphones with more powerful processors and not SBCs with the low-end CPUs. This paper introduces a CNN-ViT hybrid network called SBCFormer, which achieves high accuracy and fast computation on such low-end CPUs. The hardware constraints of these CPUs make the Transformer's attention mechanism preferable to convolution. However, using attention on low-end CPUs presents a challenge: high-resolution internal feature maps demand excessive computational resources, but reducing their resolution results in the loss of local image details. SBCFormer introduces an architectural design to address this issue. As a result, SBCFormer achieves the highest trade-off between accuracy and speed on a Raspberry Pi 4 Model B with an ARM-Cortex A72 CPU. For the first time, it achieves an ImageNet-1K top-1 accuracy of around 80% at a speed of 1.0 frame/sec on the SBC. Code is available at https://github.com/xyongLu/SBCFormer.

****

Fixing Overconfidence in Dynamic Neural Networks

Lassi Meronen, Martin Trapp, Andrea Pilzer, Le Yang, Arno Solin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2680-2690

Dynamic neural networks are a recent technique that promises a remedy for the increasing size of modern deep learning models by dynamically adapting their compu tational cost to the difficulty of the inputs. In this way, the model can adjust to a limited computational budget. However, the poor quality of uncertainty estimates in deep learning models makes it difficult to distinguish between hard and easy samples. To address this challenge, we present a computationally efficient approach for post-hoc uncertainty quantification in dynamic neural networks. We show that adequately quantifying and accounting for both aleatoric and epistemic uncertainty through a probabilistic treatment of the last layers improves the predictive performance and aids decision-making when determining the computational budget. In the experiments, we show improvements on CIFAR-100, ImageNet, and Caltech-256 in terms of accuracy, capturing uncertainty, and calibration error.

****

Multispectral Imaging for Differential Face Morphing Attack Detection: A Preliminary Study

Raghavendra Ramachandra, Sushma Venkatesh, Naser Damer, Narayan Vetrekar, R. S. Gad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6185-6193

Face morphing attack detection is emerging as an increasingly challenging problem owing to advancements in high-quality and realistic morphing attack generation. Reliable detection of morphing attacks is essential because these attacks are targeted for border control applications. This paper presents a multispectral framework for differential morphing-attack detection (D-MAD). The D-MAD methods are based on using two facial images that are captured from the ePassport (also called the reference image) and the trusted device (for example, Automatic Border Control (ABC) gates) to detect whether the face image presented in ePassport is morphed. The proposed multispectral D-MAD framework introduce a multispectral image captured as a trusted capture to acquire seven different spectral bands to detect morphing attacks. Extensive experiments were conducted on the newly created Multispectral Morphed Datasets (MSMD) with 143 unique data subjects that were captured using both visible and multispectral cameras in multiple sessions. The results indicate the superior performance of the proposed multispectral framework compared to visible images.

****

Learning Robust Deep Visual Representations From EEG Brain Recordings

Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7553-7562

Decoding the human brain has been a hallmark of neuroscientists and Artificial Intelligence researchers alike. Reconstruction of visual images from brain Electroencephalography (EEG) signals has garnered a lot of interest due to its applications in brain-computer interfacing. This study proposes a two-stage method where the first step is to obtain EEG-derived features for robust learning of deep representations and subsequently utilize the learned representation for image generation and classification. We demonstrate the generalizability of our feature extraction pipeline across three different datasets using deep-learning architectures with supervised and contrastive learning methods. We have performed the zero-shot EEG classification task to support the generalizability claim further. We observed that a subject invariant linearly separable visual representation was learned using EEG data alone in an unimodal setting that gives better k-means accuracy as compared to a joint representation learning between EEG and images. Finally, we propose a novel framework to transform unseen images into the EEG space and reconstruct them with approximation, showcasing the potential for image reconstruction from EEG signals. Our proposed image synthesis method from EEG shows 62.9% and 36.13% inception score improvement on the EEGCVPR40 and the Thoughtviz datasets, which is better than state-of-the-art performance in GAN.

****

Spiking Denoising Diffusion Probabilistic Models

Jiahang Cao, Ziqing Wang, Hanzhong Guo, Hao Cheng, Qiang Zhang, Renjing Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4912-4921

Spiking neural networks (SNNs) have ultra-low energy consumption and high biolog ical plausibility due to their binary and bio-driven nature compared with artificial neural networks (ANNs). While previous research has primarily focused on enhancing the performance of SNNs in classification tasks, the generative potential of SNNs remains relatively unexplored. In our paper, we put forward Spiking Denoising Diffusion Probabilistic Models (SDDPM), a new class of SNN-based generative models that achieve high sample quality. To fully exploit the energy efficiency of SNNs, we propose a purely Spiking U-Net architecture, which achieves comparable performance to its ANN counterpart using only 4 time steps, resulting in significantly reduced energy consumption. Extensive experimental results reveal that our approach achieves state-of-the-art on the generative tasks and substantially outperforms other SNN-based generative models, achieving up to 12x and 6x improvement on the CIFAR-10 and the CelebA datasets, respectively. Moreover, we propose a threshold-guided strategy that can further improve the performances by 2.69% in a training-free manner. The SDDPM symbolizes a significant advancement in the field of SNN generation, injecting new perspectives and potential avenues of exploration. Our code is available at https://github.com/AndyCao1125/SDDPM.

****

An Analysis of Initial Training Strategies for Exemplar-Free Class-Incremental Learning

Grégoire Petit, Michaël Soumm, Eva Feillet, Adrian Popescu, Bertrand Delezoide, David Picard, Céline Hudelot; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1837-1847

Class-Incremental Learning (CIL) aims to build classification models from data streams. At each step of the CIL process, new classes must be integrated into the model. Due to catastrophic forgetting, CIL is particularly challenging when examples from past classes cannot be stored, the case on which we focus here. To date, most approaches are based exclusively on the target dataset of the CIL process. However, the use of models pre-trained in a self-supervised way on large amounts of data has recently gained momentum. The initial model of the CIL process may only use the first batch of the target dataset, or also use pre-trained weights obtained on an auxiliary dataset. The choice between these two initial learning strategies can significantly influence the performance of the incremental learning model, but has not yet been studied in depth. Performance is also influenced by the choice of the CIL algorithm, the neural architecture, the nature of the target task, the distribution of classes in the stream and the number of examples available for learning. We conduct a comprehensive experimental study to assess the roles of these factors. We present a statistical analysis framework that quantifies the relative contribution of each factor to incremental performance. Our main finding is that the initial training strategy is the dominant factor influencing the average incremental accuracy, but that the choice of CIL algorithm is more important in preventing forgetting. Based on this analysis, we propose practical recommendations for choosing the right initial training strategy for a given incremental learning use case. These recommendations are intended to facilitate the practical deployment of incremental learning.

****

Taming Normalizing Flows

Shimon Malnick, Shai Avidan, Ohad Fried; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4644-4654

We propose an algorithm for taming Normalizing Flow models - changing the probability that the model will produce a specific image or image category. We focus on Normalizing Flows because they can calculate the exact generation probability likelihood for a given image. We demonstrate taming using models that generate human faces, a subdomain with many interesting privacy and bias considerations. Our method can be used in the context of privacy, e.g., removing a specific person from the output of a model, and also in the context of debiasing by forcing a model to output specific image categories according to a given distribution. Taming is achieved with a fast fine-tuning process without retraining from scratch, achieving the goal in a matter of minutes. We evaluate our method qualitatively and quantitatively, showing that the generation quality remains intact, while the desired changes are applied.

****

Booster-SHOT: Boosting Stacked Homography Transformations for Multiview Pedestrian Detection With Attention

Jinwoo Hwang, Philipp Benz, Pete Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 363-372

Improving multi-view aggregation is integral for multi-view pedestrian detection, which aims to obtain a bird's-eye-view pedestrian occupancy map from images captured through a set of calibrated cameras. Inspired by the success of attention modules for deep neural networks, we first propose a Homography Attention Module (HAM) which is shown to boost the performance of existing end-to-end multiview detection approaches by utilizing a novel channel gate and spatial gate. Additionally, we propose Booster-SHOT, an end-to-end convolutional approach to multiview pedestrian detection incorporating our proposed HAM as well as elements from previous approaches such as view-coherent augmentation or stacked homography transformations. Booster-SHOT achieves 92.9% and 94.2% for MODA on Wildtrack and MultiviewX respectively, outperforming the state-of-the-art by 1.4% on Wildtrack and 0.5% on MultiviewX, achieving state-of-the-art performance overall for standard evaluation metrics used in multi-view pedestrian detection.

****

ZRG: A Dataset for Multimodal 3D Residential Rooftop Understanding

Isaac Corley, Jonathan Lwowski, Peyman Najafirad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4635-4643

A crucial part of any home is the roof over our heads to protect us from the elements. In this paper we present the Zeitview Rooftop Geometry (ZRG) dataset for residential rooftop understanding. ZRG is a large-scale residential rooftop inspection dataset of over 20k properties from across the U.S. and includes high resolution aerial orthomosaics, digital surface models (DSM), colored point clouds, and 3D roof wireframe annotations. We provide an in-depth analysis and perform several experimental baselines including roof outline extraction, monocular height estimation, and planar roof structure extraction, to illustrate a few of the numerous applications unlocked by this dataset.

****

Beyond Self-Attention: Deformable Large Kernel Attention for Medical Image Segmentation

Reza Azad, Leon Niggemeier, Michael Hüttemann, Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Yury Velichko, Ulas Bagci, Dorit Merhof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1287-1297

Medical image segmentation has seen significant improvements with transformer models, which excel in grasping far-reaching contexts and global contextual information. However, the increasing computational demands of these models, proportional to the squared token count, limit their depth and resolution capabilities. Most current methods process D volumetric image data slice-by-slice (called pseudo 3D), missing crucial inter-slice information and thus reducing the model's overall performance. To address these challenges, we introduce the concept of Deformable Large Kernel Attention (D-LKA Attention), a streamlined attention mechanism employing large convolution kernels to fully appreciate volumetric context. This mechanism operates within a receptive field akin to self-attention while sidestepping the computational overhead. Additionally, our proposed attention mechanism benefits from deformable convolutions to flexibly warp the sampling grid, enabling the model to adapt appropriately to diverse data patterns. We designed both 2D and 3D adaptations of the D-LKA Attention, with the latter excelling in cross-depth data understanding. Together, these components shape our novel hierarchical Vision Transformer architecture, the D-LKA Net. Evaluations of our model against leading methods on popular medical segmentation datasets (Synapse, NIH Pancreas, and Skin lesion) demonstrate its superior performance.

****

OOD Aware Supervised Contrastive Learning

Soroush Seifi, Daniel Olmeda Reino, Nikolay Chumerin, Rahaf Aljundi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20

24, pp. 1956-1966

Out-of-Distribution (OOD) detection is a crucial problem for the safe deployment of machine learning models identifying samples that fall outside of the training distribution, i.e. in-distribution data (ID). Most OOD works focus on the classification models trained with Cross Entropy (CE) and attempt to fix its inherent issues. In this work we leverage powerful representation learned with Supervised Contrastive (SupCon) training and propose a holistic approach to learn a classifier robust to OOD data. We extend SupCon loss with two additional contrast terms. The first term pushes auxiliary OOD representations away from ID representations without imposing any constraints on similarities among auxiliary data. The second term pushes OOD features far from the existing class prototypes, while pushing ID representations closer to their corresponding class prototype. When auxiliary OOD data is not available, we propose feature mixing techniques to efficiently generate pseudo-OOD features. Our solution is simple and efficient and acts as a natural extension of the closed-set supervised contrastive representation learning. We compare against different OOD detection methods on the common benchmarks and show state-of-the-art results.

****

Meta-Learned Kernel for Blind Super-Resolution Kernel Estimation

Royson Lee, Rui Li, Stylianos Venieris, Timothy Hospedales, Ferenc Huszár, Nicholas D. Lane; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1496-1505

Recent image degradation estimation methods have enabled single-image super-resolution (SR) approaches to better upsample real-world images. Among these methods, explicit kernel estimation approaches have demonstrated unprecedented performance at handling unknown degradations. Nonetheless, a number of limitations constrain their efficacy when used by downstream SR models. Specifically, this family of methods yields i) excessive inference time due to long per-image adaptation times and ii)inferior image fidelity due to kernel mismatch. In this work, we introduce a learning-to-learn approach that meta-learns from the information contained in a distribution of images, thereby enabling significantly faster adaptation to new images with substantially improved performance in both kernel estimation and image fidelity. Specifically, we meta-train a kernel-generating GAN, named MetaKernelGAN, on a range of tasks, such that when a new image is presented, the generator starts from an informed kernel estimate and the discriminator starts with a strong capability to distinguish between patch distributions. Compared with state-of-the-art methods, our experiments show that MetaKernelGAN better estimates the magnitude and covariance of the kernel, leading to state-of-the-art blind SR results within a similar computational regime when combined with a non-blind SR model. Through supervised learning of an unsupervised learner, our method maintains the generalizability of the unsupervised learner, improves the optimization stability of kernel estimation, and hence image adaptation, and leads to a faster inference with a speedup between 14.24 to 102.1x over existing methods.

****

DDAM-PS: Diligent Domain Adaptive Mixer for Person Search

Mohammed Khaleed Almansoori, Mustansar Fiaz, Hisham Cholakkal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6688-6697

Person search (PS) is a challenging computer vision problem where the objective is to achieve joint optimization for pedestrian detection and re-Person search (PS) is a challenging computer vision problem where the objective is to achieve joint optimization for pedestrian detection and re-identification (ReID). Although previous advancements have shown promising performance in the field under fully and weakly supervised learning fashion, there exists a major gap in investigating the domain adaptation ability of PS models. In this paper, we propose a diligent domain adaptive mixer (DDAM) for person search (DDAP-PS) framework that aims to bridge a gap to improve knowledge transfer from the labeled source domain to the unlabeled target domain. Specifically, we introduce a novel DDAM module that generates moderate mixed-domain representations by combining source and targe

t domain representations. The proposed DDAM module encourages domain mixing to minimize the distance between the two extreme domains, thereby enhancing the ReID task. To achieve this, we introduce two bridge losses and a disparity loss. The objective of the two bridge losses is to guide the moderate mixed-domain representations to maintain an appropriate distance from both the source and target domain representations. The disparity loss aims to prevent the moderate mixed-domain representations from being biased towards either the source or target domains, thereby avoiding overfitting. Furthermore, we address the conflict between the two subtasks, localization and ReID, during domain adaptation. To handle this cross-task conflict, we forcefully decouple the norm-aware embedding, which aids in better learning of the moderate mixed-domain representation. We conduct experiments to validate the effectiveness of our proposed method. Our approach demonstrates favorable performance on the challenging PRW and CUHK-SYSU datasets. Our code is publicly available at https://github.com/mustansarfiaz/DDAM.
****

ArtQuest: Countering Hidden Language Biases in ArtVQA
Tibor Bleidt, Sedigheh Eslami, Gerard de Melo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7326-7335
The task of Visual Question Answering (VQA) has been studied extensively on general-domain real-world images. Transferring insights from general domain VQA to the art domain (ArtVQA) is non-trivial, as the latter requires models to identify abstract concepts, details of brushstrokes and styles of paintings in the visual data as well as possess background knowledge about art. This is exacerbated by the lack of high-quality datasets. In this work, we shed light on hidden linguistic biases in the AQUA dataset, which is the only publicly available benchmark dataset for ArtVQA. As a result, the majority of questions can be answered without consulting the visual information, making the "V" in ArtVQA rather insignificant. In order to counter this problem, we create a simple, yet practical dataset, ArtQuest, using structured information from the SemArt collection. Our dataset and the pipeline to reproduce our results are publicly available at https://github.com/bletib/artquest.
****

ISAR: A Benchmark for Single- and Few-Shot Object Instance Segmentation and Re-Identification
Nicolas Gorlo, Kenneth Blomqvist, Francesco Milano, Roland Siegwart; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4384-4396
Most object-level mapping systems in use today make use of an upstream learned object instance segmentation model. If we want to teach them about a new object or segmentation class, we need to build a large dataset and retrain the system. To build spatial AI systems that can quickly be taught about new objects, we need to effectively solve the problem of single-shot object detection, instance segmentation and re-identification. So far there is neither a method fulfilling all of these requirements in unison nor a benchmark that could be used to test such a method. Addressing this, we propose ISAR, a benchmark and baseline method for single- and few-shot object Instance Segmentation And Re-identification, in an effort to accelerate the development of algorithms that can robustly detect, segment, and re-identify objects from a single or a few sparse training examples. We provide a semi-synthetic dataset of video sequences with ground-truth semantic annotations, a standardized evaluation pipeline, and a baseline method. Our benchmark aligns with the emerging research trend of unifying Multi-Object Tracking, Video Object Segmentation, and Re-identification.
****

Textron: Weakly Supervised Multilingual Text Detection Through Data Programming
Dhruv Kudale, Badri Vishal Kasuba, Venkatapathy Subramanian, Parag Chaudhuri, Ganesh Ramakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2871-2880
Several recent deep learning (DL) based techniques perform considerably well on image-based multilingual text detection. However, their performance relies heavily on the availability and quality of training data. There are numerous types of

page-level document images consisting of information in several modalities, languages, fonts, and layouts. This makes text detection a challenging problem in the field of computer vision (CV), especially for low-resource or handwritten languages. Furthermore, there is a scarcity of word-level labeled data for text detection, especially for multilingual settings and Indian scripts that incorporate both printed and handwritten text. Conventionally, Indian script text detection requires training a DL model on plenty of labeled data, but to the best of our knowledge, no relevant datasets are available. Manual annotation of such data requires a lot of time, effort, and expertise. In order to solve this problem, we propose TEXTRON, a Data Programming-based approach, where users can plug various text detection methods into a weak supervision-based learning framework. One can view this approach to multilingual text detection as an ensemble of different CV-based techniques and DL approaches. TEXTRON can leverage the predictions of DL models pre-trained on a significant amount of language data in conjunction with CV-based methods to improve text detection in other languages. We demonstrate that TEXTRON can improve the detection performance for documents written in Indian languages, despite the absence of corresponding labeled data. Further, through extensive experimentation, we show improvement brought about by our approach over the current State-of-the-art (SOTA) models, especially for handwritten Devanagari text. Code and dataset has been made available at https://github.com/IITB-LEAP-OCR/TEXTRON

****

Sharp-NeRF: Grid-Based Fast Deblurring Neural Radiance Fields Using Sharpness Prior

Byeonghyeon Lee, Howoong Lee, Usman Ali, Eunbyung Park; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3709-3718

Neural Radiance Fields (NeRF) has shown its remarkable performance in neural rendering-based novel view synthesis. However, NeRF suffers from severe visual quality degradation when the input images have been captured under imperfect conditions, such as poor illumination, defocus blurring and lens aberrations. Especially, defocus blur is quite common in the images when they are normally captured using cameras. Although few recent studies have proposed to render sharp images of considerably high-quality, yet they still face many key challenges. In particular, those methods have employed a Multi-Layer Perceptron (MLP) based NeRF which requires tremendous computational time. To overcome these shortcomings, this paper proposes a novel technique Sharp-NeRF---a grid-based NeRF that renders clean and sharp images from the input blurry images within a half an hour training. To do so, we used several grid-based kernels to accurately model the sharpness/blurriness of the scene. The sharpness level of the pixels is computed to learn the spatially varying blur kernels. We have conducted experiments on the benchmarks consisting of blurry images and have evaluated full-reference and non-reference metrics. The qualitative and quantitative results have revealed that our approach renders the sharp novel views with vivid colors and fine details, and it has considerably faster training time than the previous works. Our code is available at https://github.com/benhenryL/SharpNeRF.

****

4K-Resolution Photo Exposure Correction at 125 FPS With ~8K Parameters

Yijie Zhou, Chao Li, Jin Liang, Tianyi Xu, Xin Liu, Jun Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1587-1597

The illumination of improperly exposed photographs has been widely corrected using deep convolutional neural networks or Transformers. Despite with promising performance, these methods usually suffer from large parameter amounts and heavy computational FLOPs on high-resolution photographs. In this paper, we propose extremely light-weight (with only 8K parameters) Multi-Scale Linear Transformation (MSLT) networks under the multi-layer perception architecture, which can process 4K-resolution sRGB images at 125 Frame-Per-Second (FPS) by a Titan RTX GPU. Specifically, the proposed MSLT networks first decompose an input image into high and low frequency layers by Laplacian pyramid techniques, and then sequentially correct different layers by pixel-adaptive linear transformation, which is implemented by efficient bilateral grid learning or 1x1 convolutions. Experiments on two benchmark datasets demonstrate the efficiency of our MSLTs against the state-of-the-arts on photo exposure correction. Extensive ablation studies validate the effectiveness of our contributions. The code is available at https://github.com/Zhou-Yijie/MSLTNet.

****

Context-Based Interpretable Spatio-Temporal Graph Convolutional Network for Human Motion Forecasting

Edgar Medina, Leyong Loh, Namrata Gurung, Kyung Hun Oh, Niels Heller; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3232-3241

Human motion prediction is still an open problem extremely important for autonomous driving and safety applications. Due to the complex spatiotemporal relation of motion sequences, this remains a challenging problem not only for movement prediction but also to perform a preliminary interpretation of the joint connections. In this work, we present a Context-based Interpretable Spatio-Temporal Graph Convolutional Network (CIST-GCN), as an efficient 3D human pose forecasting model based on GCNs that encompasses specific layers, aiding model interpretability and providing information that might be useful when analyzing motion distribution and body behavior. Our architecture extracts meaningful information from pose sequences, aggregates displacements and accelerations into the input model, and finally predicts the output displacements. Extensive experiments on Human 3.6M, AMASS, 3DPW, and ExPI datasets demonstrate that CIST-GCN outperforms previous methods in human motion prediction and robustness. Since the idea of enhancing interpretability for motion prediction has its merits, we showcase experiments towards it and provide preliminary evaluations of such insights here.

****

TPSeNCE: Towards Artifact-Free Realistic Rain Generation for Deraining and Object Detection in Rain

Shen Zheng, Changjie Lu, Srinivasa G. Narasimhan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5394-5403

Rain generation algorithms have the potential to improve the generalization of deraining methods and scene understanding in rainy conditions. However, in practice, they produce artifacts and distortions and struggle to control the amount of rain generated due to a lack of proper constraints. In this paper, we propose an unpaired image-to-image translation framework for generating realistic rainy images. We first introduce a Triangular Probability Similarity (TPS) constraint to guide the generated images toward clear and rainy images in the discriminator manifold, thereby minimizing artifacts and distortions during rain generation. Unlike conventional contrastive learning approaches, which indiscriminately push negative samples away from the anchors, we propose a Semantic Noise Contrastive Estimation (SeNCE) strategy and reassess the pushing force of negative samples based on the semantic similarity between the clear and the rainy images and the feature similarity between the anchor and the negative samples. Experiments demonstrate realistic rain generation with minimal artifacts and distortions, which benefits image deraining and object detection in rain. Furthermore, the method can be used to generate realistic snowy and night images, underscoring its potential for broader applicability. Code is available at https://github.com/ShenZheng2000/TPSeNCE.

****

Robust Category-Level 3D Pose Estimation From Diffusion-Enhanced Synthetic Data

Jiahao Yang, Wufei Ma, Angtian Wang, Xiaoding Yuan, Alan Yuille, Adam Kortylewski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3446-3455

Obtaining accurate 3D object poses is vital for numerous computer vision applications, such as 3D reconstruction and scene understanding. However, annotating real-world objects is time-consuming and challenging. While synthetically generated training data is a viable alternative, the domain shift between real and synthetic data is a significant challenge. In this work, we aim to narrow the perform ance gap between models trained on synthetic data and fully supervised models trained on a large amount of real data. We achieve this by approaching the problem from two perspectives: 1) We introduce P3D-Diffusion, a new synthetic dataset with accurate 3D annotations generated with a graphics-guided diffusion model. 2) We propose Cross-domain 3D Consistency, CC3D, for unsupervised domain adaptation of neural mesh models. In particular, we exploit the spatial relationships between features on the mesh surface and a contrastive learning scheme to guide the domain adaptation process. Combined, these two approaches enable our models to perform competitively with state-of-the-art models using only 10% of the respective real training images, while outperforming the SOTA model by a wide margin using only 50% of the real training data. By encouraging the diversity of synthetic data and generating the images with an OOD-aware manner, our model further demonstrates robust generalization to out-of-distribution scenarios despite being trained with minimal real data.

****

Vision Transformer for Multispectral Satellite Imagery: Advancing Landcover Classification

Ryan Rad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8176-8183

Climate change is a global issue with significant impacts on ecosystems and human populations. Accurately classifying land cover from multi-spectral satellite imagery plays a crucial role in understanding the Earth's changing landscape and its implications for environmental processes. However, traditional methods struggle with challenges like limited data availability and capturing complex spatial-spectral relationships. Vision Transformers have emerged as a promising alternative to convolutional neural networks (CNN architectures), harnessing the power of self-attention mechanisms to capture global and long-range dependencies. However, their application to multi-spectral images is still limited. In this paper, we propose a novel Vision Transformer designed for multi-spectral satellite image datasets of limited size to perform reliable land cover identification with forty-four classes. We conduct extensive experiments on a curated dataset, simulating scenarios with limited data availability, and compare our approach to alternative architectures. The results demonstrate the potential of our Vision Transformer-based method in achieving accurate land cover classification, contributing to improving climate change modeling and environmental understanding.

****

ENTED: Enhanced Neural Texture Extraction and Distribution for Reference-Based Blind Face Restoration

Yuen-Fui Lau, Tianjia Zhang, Zhefan Rao, Qifeng Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5162-5171

We present ENTED, a new framework for blind face restoration that aims to restore high-quality and realistic portrait images. Our method involves repairing a single degraded input image using a high-quality reference image. We utilize a texture extraction and distribution framework to transfer high-quality texture features between the degraded input and reference image. However, the StyleGAN-like architecture in our framework requires high-quality latent codes to generate realistic images. The latent code extracted from the degraded input image often contains corrupted features, making it difficult to align the semantic information from the input with the high-quality textures from the reference. To overcome this challenge, we employ two special techniques. The first technique, inspired by vector quantization, replaces corrupted semantic features with high-quality code words. The second technique generates style codes that carry photorealistic texture information from a more informative latent space developed using the high-quality features in the reference image's manifold. Extensive experiments conducted on synthetic and real-world datasets demonstrate that our method produces results with more realistic contextual details and outperforms state-of-the-art methods. A thorough ablation study confirms the effectiveness of each proposed module.

****

A Sequential Learning-Based Approach for Monocular Human Performance Capture
Jianchun Chen, Jayakorn Vongkulbhisal, Fernando De la Torre Frade; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3514-3523
Human performance capture from RGB videos in unconstrained environments has become very popular for applications that require generating virtual avatars or digital actors. SOTA methods use neural network (NN) techniques to estimate the shape directly from photos, yielding a simplified model of the human body. While effective, NN techniques frequently fail under challenging poses and do not preserve temporal consistency. On the other hand, optimization-based methods like shape-from-silhouette can produce more precise reconstruction; however, they typically require a good initialization and are computationally more intensive than NN. To address issues of previous methods, this work proposes a learning-based approach for optimizing fine-grained shape representation (e.g., clothes, wrinkles) from a monocular RGB video. Our main idea is to sequentially recover different shape details (e.g., average shape, clothing, wrinkles) using separate neural networks. At each level, our network takes the sparse/noisy gradients of body mesh vertices w.r.t the shape, and predicts dense gradients to update the body shape. Despite being trained on synthetic data, these networks have surprisingly good generalization to real images. Experimental validation shows that our approach outperforms NN approaches in recovering shape details while also being an order of magnitude faster than optimization-based methods and robust across varied poses and novel views.
****
VCISR: Blind Single Image Super-Resolution With Video Compression Synthetic Data
Boyang Wang, Bowen Liu, Shiyu Liu, Fengyu Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4302-4312
In the blind single image super-resolution (SISR) task, existing works have been successful in restoring image-level unknown degradations. However, when a single video frame becomes the input, these works usually fail to address degradations caused by video compression, such as mosquito noise, ringing, blockiness, and staircase noise. In this work, we for the first time, present a video compression-based degradation model to synthesize low-resolution image data in the blind SISR task. Our proposed image synthesizing method is widely applicable to existing image datasets, so that a single degraded image can contain distortions caused by the lossy video compression algorithms. This overcomes the leak of feature diversity in video data and thus retains the training efficiency. By introducing video coding artifacts to SISR degradation models, neural networks can super-resolve images with the ability to restore video compression degradations, and achieve better results on restoring generic distortions caused by image compression as well. Our proposed approach achieves superior performance in SOTA no-reference Image Quality Assessment, and shows better visual quality on various datasets. In addition, we evaluate the SISR neural network trained with our degradation model on video super-resolution (VSR) datasets. Compared to architectures specifically designed for the VSR purpose, our method exhibits similar or better performance, evidencing that the presented strategy on infusing video-based degradation is generalizable to address more complicated compression artifacts even without temporal cues. The code is available at https://github.com/Kiteretsu77/VCISR-official.
****
Synthesizing Coherent Story With Auto-Regressive Latent Diffusion Models
Xichen Pan, Pengda Qin, Yuhong Li, Hui Xue, Wenhu Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2920-2930
Conditioned diffusion models have demonstrated state-of-the-art text-to-image synthesis capacity. Recently, most works focus on synthesizing independent images; While for real-world applications, it is common and necessary to generate a series of coherent images for story-stelling. In this work, we mainly focus on story visualization and continuation tasks and propose AR-LDM, a latent diffusion model auto-regressively conditioned on history captions and generated images. More

over, AR-LDM can generalize to new characters through adaptation. To our best knowledge, this is the first work successfully leveraging diffusion models for coherent visual story synthesizing. It also extends the text-conditioned method to multimodal conditioning. Quantitative results show that AR-LDM achieves SoTA FID scores on PororoSV, FlintstonesSV, and the adopted challenging dataset VIST containing natural images. Large-scale human evaluations show that AR-LDM has superior performance in terms of quality, relevance, and consistency.
****

Text-to-Image Editing by Image Information Removal
Zhongping Zhang, Jian Zheng, Zhiyuan Fang, Bryan A. Plummer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5232-5241

Diffusion models have demonstrated impressive performance in text-guided image generation. Current methods that leverage the knowledge of these models for image editing either fine-tune them using the input image (e.g., Imagic) or incorporate structure information as additional constraints (e.g., ControlNet). However, fine-tuning large-scale diffusion models on a single image can lead to severe overfitting issues and lengthy inference time. Information leakage from pretrained models also make it challenging to preserve image content not related to the text input. Additionally, methods that incorporate structural guidance (e.g., edge maps, semantic maps, keypoints) find retaining attributes like colors and textures difficult. Using the input image as a control could mitigate these issues, but since these models are trained via reconstruction, a model can simply hide information about the original image when encoding it to perfectly reconstruct the image without learning the editing task. To address these challenges, we propose a text-to-image editing model with an Image Information Removal module (IIR) that selectively erases color-related and texture-related information from the original image, allowing us to better preserve the text-irrelevant content and avoid issues arising from information hiding. Our experiments on CUB, Outdoor Scenes, and COCO reports our approach achieves the best editability-fidelity trade-off results. In addition, a user study on COCO shows that our edited images are preferred 35% more often than prior work.
****

Self-Annotated 3D Geometric Learning for Smeared Points Removal
Miaowei Wang, Daniel Morris; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3494-3503

There has been significant progress in improving the accuracy and quality of consumer-level dense depth sensors. Nevertheless, there remains a common depth pixel artifact which we call smeared points. These are points not on any 3D surface and typically occur as interpolations between foreground and background objects. As they cause fictitious surfaces, these points have the potential to harm applications dependent on the depth maps. Statistical outlier removal methods fare poorly in removing these points as they tend also to remove actual surface points. Trained network-based point removal faces difficulty in obtaining sufficient annotated data. To address this, we propose a fully self-annotated method to train a smeared point removal classifier. Our approach relies on gathering 3D geometric evidence from multiple perspectives to automatically detect and annotate smeared points and valid points. To validate the effectiveness of our method, we present a new benchmark dataset: the Real Azure-Kinect dataset. Experimental results and ablation studies show that our method outperforms traditional filters and other self-annotated methods. Our work is publicly available at https://github.com/wangmiaowei/wacv2024_smearedremover.git.
****

Deep Metric Learning With Chance Constraints
Yeti Z. Gürbüz, O■ul Can, Aydin Alatan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 543-553

Deep metric learning (DML) aims to minimize empirical expected loss of the pairwise intra-/inter- class proximity violations in the embedding space. We relate DML to feasibility problem of finite chance constraints. We show that minimizer of proxy-based DML satisfies certain chance constraints, and that the worst case generalization performance of the proxy-based methods can be characterized by the radius of the smallest ball around a class proxy to cover the entire domain of the corresponding class samples, suggesting multiple proxies per class helps performance. To provide a scalable algorithm as well as exploiting more proxies, we consider the chance constraints implied by the minimizers of proxy-based DML instances and reformulate DML as finding a feasible point in intersection of such constraints, resulting in a problem to be approximately solved by iterative projections. Simply put, we repeatedly train a regularized proxy-based loss and re-initialize the proxies with the embeddings of the deliberately selected new samples. We applied our method with 4 well-accepted DML losses and show the effectiveness with extensive evaluations on 4 popular DML benchmarks. Code is available at: https://github.com/yetigurbuz/ccp-dml

****

CrashCar101: Procedural Generation for Damage Assessment

Jens Parslov, Erik Riise, Dim P. Papadopoulos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4624-4634

In this paper, we are interested in addressing the problem of damage assessment for vehicles, such as cars. This task requires not only detecting the location and the extent of the damage but also identifying the damaged part. To train a computer vision system for the semantic part and damage segmentation in images, we need to manually annotate images with costly pixel annotations for both part categories and damage types. To overcome this need, we propose to use synthetic data to train these models. Synthetic data can provide samples with high variability, pixel-accurate annotations, and arbitrarily large training sets without any human intervention. We propose a procedural generation pipeline that damages 3D car models and we obtain synthetic 2D images of damaged cars paired with pixel-accurate annotations for part and damage categories. To validate our idea, we execute our pipeline and render our CrashCar101 dataset. We run experiments on three real datasets for the tasks of part and damage segmentation. For part segmentation, we show that the segmentation models trained on a combination of real data and our synthetic data outperform all models trained only on real data. For damage segmentation, we show the sim2real transfer ability of CrashCar101.

****

Towards Domain-Aware Knowledge Distillation for Continual Model Generalization

Nikhil Reddy, Mahsa Baktashmotlagh, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 696-707

Generalization on unseen domains is critical for Deep Neural Networks (DNNs) to perform well in real-world applications such as autonomous navigation. However, catastrophic forgetting limit the ability of domain generalization and unsupervised domain adaption approaches to adapt to constantly changing target domains. To overcome these challenges, We propose DoSe framework, a Domain-aware Self-Distillation method based on batch normalization prototypes to facilitate continual model generalization across varying target domains. Specifically, we enforce the consistency of batch normalization statistics between two batches of images sampled from the same target domain distribution between the student and teacher models. To alleviate catastrophic forgetting, we introduce a novel exemplar-based replay buffer to identify difficult samples for the model to retain the knowledge. Specifically, we demonstrate that identifying difficult samples and updating the model periodically using them can help in preserving knowledge learned from previously seen domains. We conduct extensive experiments on two real-world datasets ACDC, C-Driving, and one synthetic dataset SHIFT to verify the efficiency of the proposed DoSe framework. On ACDC, our method outperforms existing SOTA in Domain Generalization, Unsupervised Domain Adaptation, and Daytime settings by 26%, 14%, and 70% respectively.

****

SCoRD: Subject-Conditional Relation Detection With Text-Augmented Data

Ziyan Yang, Kushal Kafle, Zhe Lin, Scott Cohen, Zhihong Ding, Vicente Ordonez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5731-5741

We propose Subject-Conditional Relation Detection SCoRD, where conditioned on an

input subject, the goal is to predict all its relations to other objects in a scene along with their locations. Based on the Open Images dataset, we propose a challenging OIv6-SCoRD benchmark such that the training and testing splits have a distribution shift in terms of the occurrence statistics of <subject, relation, object> triplets. To solve this problem, we propose an auto-regressive model that given a subject, it predicts its relations, objects, and object locations by casting this output as a sequence of tokens. First, we show that previous scene-graph prediction methods fail to produce as exhaustive an enumeration of relation-object pairs when conditioned on a subject on this benchmark. Particularly, we obtain a recall@3 of 83.8% for our relation-object predictions compared to the 49.75% obtained by a recent scene graph detector. Then, we show improved generalization on both relation-object and object-box predictions by leveraging during training relation-object pairs obtained automatically from textual captions and for which no object-box annotations are available. Particularly, for <subject, relation, object> triplets for which no object locations are available during training, we are able to obtain a recall@3 of 33.80% for relation-object pairs and 26.75% for their box locations.
****
THInImg: Cross-Modal Steganography for Presenting Talking Heads in Images
Lin Zhao, Hongxuan Li, Xuefei Ning, Xinru Jiang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5553-5562
Cross-modal Steganography is the practice of concealing secret signals in publicly available cover signals (distinct from the modality of the secret signals) unobtrusively. While previous approaches primarily concentrated on concealing a relatively small amount of information, we propose THInImg, which manages to hide lengthy audio data (and subsequently decode talking head video) inside an identity image by leveraging the properties of human face, which can be effectively utilized for covert communication, transmission and copyright protection. THInImg consists of two parts: the encoder and decoder. Inside the encoder-decoder pipeline, we introduce a novel architecture that substantially increase the capacity of hiding audio in images. Moreover, our framework can be extended to iteratively hide multiple audio clips into an identity image, offering multiple levels of control over permissions. We conduct extensive experiments to prove the effectiveness of our method, demonstrating that THInImg can present up to 80 seconds of high quality talking-head video (including audio) in an identity image with 160x160 resolution.
****
Causal Analysis for Robust Interpretability of Neural Networks
Ola Ahmad, Nicolas Béreux, Loïc Baret, Vahid Hashemi, Freddy Lecue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4685-4694
Interpreting the inner function of neural networks is crucial for the trustworthy development and deployment of these black-box models. Prior interpretability methods focus on correlation-based measures to attribute model decisions to individual examples. However, these measures are susceptible to noise and spurious correlations encoded in the model during the training phase (e.g., biased inputs, model overfitting, or misspecification). Moreover, this process has proven to result in noisy and unstable attributions that prevent any transparent understanding of the model's behavior. In this paper, we develop a robust interventional-based method grounded by causal analysis to capture cause-effect mechanisms in pre-trained neural networks and their relation to the prediction. Our novel approach relies on path interventions to infer the causal mechanisms within hidden layers and isolate relevant and necessary information (to model prediction), avoiding noisy ones. The result is task-specific causal explanatory graphs that can audit model behavior and express the actual causes underlying its performance. We apply our method to vision models trained on classification tasks. On image classification tasks, we provide extensive quantitative experiments to show that our approach can capture more stable and faithful explanations than standard attribution-based methods. Furthermore, the underlying causal graphs express the neural interactions in the model, making it a valuable tool in other applications (e.g

., model repair).

****

TransFed: A Way To Epitomize Focal Modulation Using Transformer-Based Federated Learning

Tajamul Ashraf, Fuzayil Bin Afzal Mir, Iqra Altaf Gillani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 554-563

Federated learning has emerged as a promising paradigm for collaborative machine learning, enabling multiple clients to train a model while preserving data privacy jointly. Tailored federated learning takes this concept further by accommodating client heterogeneity and facilitating the learning of personalized models. While the utilization of transformers within federated learning has attracted significant interest, there remains a need to investigate the effects of federated learning algorithms on the latest focal modulation-based transformers. In this paper, we investigate this relationship and uncover the detrimental effects of federated averaging (FedAvg) algorithms on Focal Modulation, particularly in scenarios with heterogeneous data. To address this challenge, we propose TransFed, a novel transformer-based federated learning framework that not only aggregates model parameters but also learns tailored Focal Modulation for each client. Instead of employing a conventional customization mechanism that maintains client-specific focal modulation layers locally, we introduce a learn-to-tailor approach that fosters client collaboration, enhancing scalability and adaptation in TransFed. Our method incorporates a hyper network on the server, responsible for learning personalized projection matrices for the focal modulation layers. This enables the generation of client-specific keys, values, and queries. Furthermore, we provide an analysis of adaptation bounds for TransFed using the learn-to-customize mechanism. Through intensive experiments on datasets related to pneumonia classification, we demonstrate that TransFed, in combination with the learn-to-tailor approach, achieves superior performance in scenarios with non-IID data distributions, surpassing existing methods. Overall, TransFed paves the way for leveraging focal Modulation in federated learning, advancing the capabilities of focal modulated transformer models in decentralized environments.

****

Natural Light Can Also Be Dangerous: Traffic Sign Misinterpretation Under Adversarial Natural Light Attacks

Teng-Fang Hsiao, Bo-Lun Huang, Zi-Xiang Ni, Yan-Ting Lin, Hong-Han Shuai, Yung-Hui Li, Wen-Huang Cheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3915-3924

Common illumination sources like sunlight or artificial light may introduce hidden vulnerabilities to AI systems. Our paper delves into these potential threats, offering a novel approach to simulate varying light conditions, including sunlight, headlights, and flashlight illuminations. Moreover, unlike typical physical adversarial attacks requiring conspicuous alterations, our method utilizes a model-agnostic black-box attack integrated with the Zeroth Order Optimization (ZOO) algorithm to identify deceptive patterns in a physically-applicable space. Consequently, attackers can recreate these simulated conditions, deceiving machine learning models with seemingly natural light. Empirical results demonstrate the efficacy of our method, misleading models trained on the GTSRB and LISA datasets under natural-like physical environments with an attack success rate exceeding 70% across all digital datasets, and remaining effective against all evaluated real-world traffic signs. Importantly, after adversarial training using samples generated from our approach, models showcase enhanced robustness, underscoring the dual value of our work in both identifying and mitigating potential threats.

****

PAIR: Perception Aided Image Restoration for Natural Driving Conditions

Pranjay Shyam, HyunJin Yoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7459-7470

We present a two-stage mechanism for generic image restoration in natural driving conditions, where multiple non-linear degradations simultaneously impact perception for humans and driving assistance systems. Our approach overcomes the limi tations of utilizing a single neural network that incurs excessive computational overhead and yields sub-optimal recovery. The proposed first stage comprises computationally inexpensive image processing operations applied at a patch level using a lightweight convolutional neural network (CNN) that determines their intensity of operation. This patch size is guided by the receptive field of the CNN, allowing for dynamic restoration of non-linear and non-homogeneous degradation profiles. The second stage leverages a lightweight end-to-end neural network functioning as an inpainting network. It identifies inadequately restored regions and leverages global semantic and structural information to fill the affected areas. This approach enhances the restoration process by considering the entire image and addresses the remainder of localized deficiencies. In addition, we integrate dense perception tasks such as semantic and depth estimation during the optimization cycle to ensure restored images that are perceptually pleasing and conducive for downstream perception tasks. Since datasets covering diverse degradation scenarios for high- and low-level perception tasks are lacking, we utilize a synthetic data augmentation technique to generate non-homogeneous non-linear degradation profiles. Experiments on images captured in adverse weather conditions demonstrate the efficacy of our approach, yielding higher perceptual quality in restored images and improved performance in downstream perception tasks under adverse driving conditions. Importantly, our method offers computational efficiency compared to end-to-end image restoration algorithms, making it suitable for real-time applications.

****

RecycleNet: Latent Feature Recycling Leads to Iterative Decision Refinement

Gregor Köhler, Tassilo Wald, Constantin Ulrich, David Zimmerer, Paul F. Jäger, Jörg K.H. Franke, Simon Kohl, Fabian Isensee, Klaus H. Maier-Hein; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 810-818

Despite the remarkable success of deep learning systems over the last decade, a key difference still remains between neural network and human decision-making: As humans, we can not only form a decision on the spot, but also ponder, revisiting an initial guess from different angles, distilling relevant information, arriving at a better decision. Here, we propose RecycleNet, a latent feature recycling method, instilling the pondering capability for neural networks to refine initial decisions over a number of recycling steps, where outputs are fed back into earlier network layers in an iterative fashion. This approach makes minimal assumptions about the neural network architecture and thus can be implemented in a wide variety of contexts. Using medical image segmentation as the evaluation environment, we show that latent feature recycling enables the network to iteratively refine initial predictions even beyond the iterations seen during training, converging towards an improved decision. We evaluate this across a variety of segmentation benchmarks and show consistent improvements even compared with top-performing segmentation methods. This allows trading increased computation time for improved performance, which can be beneficial, especially for safety-critical applications.

****

CamoFocus: Enhancing Camouflage Object Detection With Split-Feature Focal Modulation and Context Refinement

Abbas Khan, Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, Giulia De Masi, Fakhri Karray; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1434-1443

Camouflage Object Detection (COD) involves the challenge of isolating a target object from a visually similar background, presenting a formidable challenge for learning algorithms. Drawing inspiration from state-of-the-art (SOTA) Focal Modulation Networks, our objective is to proficiently modulate the foreground and background components, thereby capturing the distinct features of each. We introduce a Feature Split and Modulation (FSM) module to attain this goal. This module efficiently separates the object from the background by utilizing foreground and background modulators guided by a supervisory mask. For enhanced feature refinement, we propose a Context Refinement Module (CRM), which considers features acq uired from FSM across various spatial scales, leading to comprehensive enrichment and highly accurate prediction maps. Through extensive experimentation, we showcase the superiority of CamoFocus over recent SOTA COD methods. Our evaluations encompass diverse benchmark datasets, including CAMO, COD10K, CHAMELEON, and NC4K. The findings underscore the potential and significance of the proposed CamoFocus model and establish its efficacy in addressing the critical challenges of camouflage object detection.

****

Scene Text Image Super-Resolution Based on Text-Conditional Diffusion Models

Chihiro Noguchi, Shun Fukuda, Masao Yamanaka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1485-1495

Scene Text Image Super-resolution (STISR) has recently achieved great success as a preprocessing method for scene text recognition. STISR aims to transform blurred and noisy low-resolution (LR) text images in real-world settings into clear high-resolution (HR) text images suitable for scene text recognition. In this study, we leverage text-conditional diffusion models (DMs), known for their impressive text-to-image synthesis capabilities, for STISR tasks. Our experimental results revealed that text-conditional DMs notably surpass existing STISR methods. Especially when texts from LR text images are given as input, the text-conditional DMs are able to produce superior quality super-resolution text images. Utilizing this capability, we propose a novel framework for synthesizing LR-HR paired text image datasets. This framework consists of three specialized text-conditional DMs, each dedicated to text image synthesis, super-resolution, and image degradation. These three modules are vital for synthesizing distinct LR and HR paired images, which are more suitable for training STISR methods. Our experiments confirmed that these synthesized image pairs significantly enhance the performance of STISR methods in the TextZoom evaluation.

****

Domain Adaptive 3D Shape Retrieval From Monocular Images

Harsh Pal, Ritwik Khandelwal, Shivam Pande, Biplab Banerjee, Srikrishna Karanam; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3192-3201

In this work, we address the novel and challenging problem of domain adaptive 3D shape retrieval from single 2D images (DA-IBSR). While the existing image-based 3D shape retrieval (IBSR) problem focuses on modality alignment for retrieving a matchable 3D shape from a shape repository given a 2D image query, it does not consider any distribution shift between the training and testing image-shape pairs, making the performance of off-the-shelves IBSR methods subpar. In contrast, the proposed DA-IBSR addresses the non-trivial problem of modality shift as well distribution shift across training and test sets. To address these issues, we propose an end-to-end trainable model called DAIS-NET. Our objective is to align the images and shapes separately from both domains while simultaneously learn a shared embedding space for the 2D and 3D modalities. The former problem is addressed by separately employing maximum mean discrepancy loss across the 2D images and 3D shapes of the two domains. To address the modality alignment, we incorporate the notion of negative sample mining and employ triplet loss to bridge the gap between positive 2D-3D pairs (of same class) and increase the separation between negative 2D-3D pairs (of different class). Additionally, we employ an entropy minimization strategy to align the unlabeled target domain data in the semantic space. To evaluate our proposed approach, we define the experimental setting of DA-IBSR on the following benchmarks: SHREC'14 <-> Pix3D and ShapeNet <-> SHREC'14. Considering the novelty of the problem statement, we have demonstrated that the issue of domain gap is prevalent by comparing our method with the existing literature. Additionally, through extensive evaluations, we demonstrate the capability of DAIS-NET to successfully mitigate this domain gap in image based 3D shape retrieval.

****

Learning Quality Labels for Robust Image Classification

Xiaosong Wang, Ziyue Xu, Dong Yang, Leo Tam, Holger Roth, Daguang Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2

024, pp. 1103-1112

Current deep learning paradigms largely benefit from the tremendous amount of annotated data. However, the quality of the annotations often varies among labelers. Multi-observer studies have been conducted to examine the annotation variances (by labeling the same data multiple times) and their effects on critical applications like medical image analysis. In this paper, we demonstrate how multiple sets of annotations (either hand-labeled or algorithm-generated) can be utilized together and mutually benefit the learning of classification tasks. The concept of learning-to-vote is introduced to sample quality label sets for each data entry on-the-fly during the training. Specifically, a meta-training-based label-sampling module is designed to achieve refined labels (weighted sum of attended ones) that benefit the model learning the most through additional back-propagations. We apply the learning-to-vote scheme on the classification task of a synthetic noisy CIFAR-10 to prove the concept and then demonstrate superior results (3-5% increase on average in multiple disease classification AUCs) on the chest x-ray images from a hospital-scale dataset (MIMIC-CXR) and hand-labeled dataset (OpenI) in comparison to regular training paradigms.

****

LibreFace: An Open-Source Toolkit for Deep Facial Expression Analysis

Di Chang, Yufeng Yin, Zongjian Li, Minh Tran, Mohammad Soleymani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8205-8215

Facial expression analysis is an important tool for human-computer interaction. In this paper, we introduce LibreFace, an open-source toolkit for facial expression analysis. This open-source toolbox offers real-time and offline analysis of facial behavior through deep learning models, including facial action unit (AU) detection, AU intensity estimation, and facial expression recognition. To accomplish this, we employ several techniques, including the utilization of a large-scale pre-trained network, feature-wise knowledge distillation, and task-specific fine-tuning. These approaches are designed to effectively and accurately analyze facial expressions by leveraging visual information, thereby facilitating the implementation of real-time interactive applications. In terms of Action Unit (AU) intensity estimation, we achieve a Pearson Correlation Coefficient (PCC) of 0.63 on DISFA, which is 7% higher than the performance of OpenFace 2.0 while maintaining highly-efficient inference that runs two times faster than OpenFace 2.0. Despite being compact, our model also demonstrates competitive performance to state-of-the-art facial expression analysis methods on AffecNet, FFHQ, and RAF-DB.

****

SCUNet++: Swin-UNet and CNN Bottleneck Hybrid Architecture With Multi-Fusion Dense Skip Connection for Pulmonary Embolism CT Image Segmentation

Yifei Chen, Binfeng Zou, Zhaoxin Guo, Yiyu Huang, Yifan Huang, Feiwei Qin, Qinhai Li, Changmiao Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7759-7767

Pulmonary embolism (PE) is a prevalent lung disease that can lead to right ventricular hypertrophy and failure in severe cases, ranking second in severity only to myocardial infarction and sudden death. Pulmonary artery CT angiography (CTPA) is a widely used diagnostic method for PE. However, PE detection presents challenges in clinical practice due to limitations in imaging technology. CTPA can produce noises similar to PE, making confirmation of its presence time-consuming and prone to overdiagnosis. Nevertheless, the traditional segmentation method of PE can not fully consider the hierarchical structure of features, local and global spatial features of PE CT images. In this paper, we propose an automatic PE segmentation method called SCUNet++ (Swin Conv UNet++). This method incorporates multiple fusion dense skip connections between the encoder and decoder, utilizing the Swin Transformer as the encoder. And fuses features of different scales in the decoder subnetwork to compensate for spatial information loss caused by the inevitable downsampling in Swin-UNet or other state-of-the-art methods, effectively solving the above problem. We provide a theoretical analysis of this method in detail and validate it on publicly available PE CT image datasets FUMPE and CAD-PE. The experimental results indicate that our proposed method achieved a D ice similarity coefficient (DSC) of 83.47% and a Hausdorff distance 95th percentile (HD95) of 3.83 on the FUMPE dataset, as well as a DSC of 83.42% and an HD95 of 5.10 on the CAD-PE dataset. These findings demonstrate that our method exhibits strong performance in PE segmentation tasks, potentially enhancing the accuracy of automatic segmentation of PE and providing a powerful diagnostic tool for clinical physicians. Our source code and new FUMPE dataset are available at https://github.com/JustlfC03/SCUNet-plusplus.

****

Attention Modules Improve Image-Level Anomaly Detection for Industrial Inspection: A DifferNet Case Study

André Luiz Vieira e Silva, Francisco Simões, Danny Kowerko, Tobias Schlosser, Felipe Battisti, Veronica Teichrieb; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8246-8255

Within (semi-)automated visual industrial inspection, learning-based approaches for assessing visual defects, including deep neural networks, enable the processing of otherwise small defect patterns in pixel size on high-resolution imagery. The emergence of these often rarely occurring defect patterns explains the general need for labeled data corpora. To alleviate this issue and advance the current state of the art in unsupervised visual inspection, this work proposes a DifferNet-based solution enhanced with attention modules: AttentDifferNet. It improves image-level detection and classification capabilities on three visual anomaly detection datasets for industrial inspection: InsPLAD-fault, MVTec AD, and Semiconductor Wafer. In comparison to the state of the art, AttentDifferNet achieves improved results, which are, in turn, highlighted throughout our quali-quantitative study. Our quantitative evaluation shows an average improvement - compared to DifferNet - of 1.77 +- 0.25 percentage points in overall AUROC considering all three datasets, reaching SOTA results in InsPLAD-fault, an industrial inspection in-the-wild dataset. As our variants to AttentDifferNet show great prospects in the context of currently investigated approaches, a baseline is formulated, emphasizing the importance of attention for industrial anomaly detection both in the wild and in controlled environments.

****

Indoor Visual Localization Using Point and Line Correspondences in Dense Colored Point Cloud

Yuya Matsumoto, Gaku Nakano, Kazumine Ogura; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3616-3625

We propose a novel pipeline called Loc-PL that uses both points and lines for indoor visual localization in dense colored point cloud. Loc-PL utilizes the spatially complementary relationship between points and lines to address challenging indoor issues. There are two successive camera pose estimation modules. The first improves robustness against repetitive patterns by considering the geometric consistency of points and lines. The second utilizes points and lines to refine poses by Perspective-m-Point-n-Line (PmPnL) and circumvents unstable localization due to locally concentrated matches caused by less-textured environments. The modules use different schemes to obtain line correspondences; the first finds line matches using RANSAC, which is effective for image pairs with large viewpoint gaps, and the second utilizes rendered images from dense point cloud to get them by feature line matching. In addition, we develop a simple but effective module for evaluating the correctness of camera poses using matched point distances across two images. The experimental results on a large dataset, InLoc, show that Loc-PL achieves the state-of-the-art in four out of six scores.

****

RGB-D Mapping and Tracking in a Plenoxel Radiance Field

Andreas L. Teigen, Yeonsoo Park, Annette Stahl, Rudolf Mester; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3342-3351

The widespread adoption of Neural Radiance Fields (NeRFs) have ensured significant advances in the domain of novel view synthesis in recent years. These models capture a volumetric radiance field of a scene, creating highly convincing, dense, photorealistic models through the use of simple, differentiable rendering equ

ations. Despite their popularity, these algorithms suffer from severe ambiguities in visual data inherent to the RGB sensor, which means that although images generated with view synthesis can visually appear very believable, the underlying 3D model will often be wrong. This considerably limits the usefulness of these models in practical applications like Robotics and Extended Reality (XR), where an accurate dense 3D reconstruction otherwise would be of significant value. In this paper, we present the vital differences between view synthesis models and 3D reconstruction models. We also comment on why a depth sensor is essential for modeling accurate geometry in general outward-facing scenes using the current paradigm of novel view synthesis methods. Focusing on the structure-from-motion task, we practically demonstrate this need by extending the Plenoxel radiance field model: Presenting an analytical differential approach for dense mapping and tracking with radiance fields based on RGB-D data without a neural network. Our method achieves state-of-the-art results in both mapping and tracking tasks, while also being faster than competing neural network-based approaches. The code is available at: https://github.com/ysus33/RGB-D_Plenoxel_Mapping_Tracking.git
****

An Empirical Investigation Into Benchmarking Model Multiplicity for Trustworthy Machine Learning: A Case Study on Image Classification

Prakhar Ganesh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4488-4497

Deep learning models have proven to be highly successful. Yet, their over-parameterization gives rise to model multiplicity, a phenomenon in which multiple models achieve similar performance but exhibit distinct underlying behaviours. This multiplicity presents a significant challenge and necessitates additional specifications in model selection to prevent unexpected failures during deployment. While prior studies have examined these concerns, they focus on individual metrics in isolation, making it difficult to obtain a comprehensive view of multiplicity in trustworthy machine learning. Our work stands out by offering a one-stop empirical benchmark of multiplicity across various dimensions of model design and its impact on a diverse set of trustworthy metrics. In this work, we establish a consistent language for studying model multiplicity by translating several trustworthy metrics into accuracy under appropriate interventions. We also develop a framework, which we call multiplicity sheets, to benchmark multiplicity in various scenarios. We demonstrate the advantages of our setup through a case study in image classification and provide actionable insights into the impact and trends of different hyperparameters on model multiplicity. Finally, we show that multiplicity persists in deep learning models even after enforcing additional specifications during model selection, highlighting the severity of over-parameterization. The concerns of under-specification thus remain, and we seek to promote a more comprehensive discussion of multiplicity in trustworthy machine learning.
****

Pixel-Grounded Prototypical Part Networks

Zachariah Carmichael, Suhas Lohit, Anoop Cherian, Michael J. Jones, Walter J. Scheirer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4768-4779

Prototypical part neural networks (ProtoPartNNs), namely ProtoPNet and its derivatives, are an intrinsically interpretable approach to machine learning. Their prototype learning scheme enables intuitive explanations of the form, this (prototype) looks like that (testing image patch). But, does this actually look like that? In this work, we delve into why object part localization and associated heat maps in past work are misleading. Rather than localizing to object parts, existing ProtoPartNNs localize to the entire image, contrary to generated explanatory visualizations. We argue that detraction from these underlying issues is due to the alluring nature of visualizations and an over-reliance on intuition. To alleviate these issues, we devise new receptive field-based architectural constraints for meaningful localization and a principled pixel space mapping for ProtoPartNNs. To improve interpretability, we propose additional architectural improvements, including a simplified classification head. We also make additional corrections to ProtoPNet and its derivatives, such as the use of a validation set, rat her than a test set, to evaluate generalization during training. Our approach, PixPNet (Pixel-grounded Prototypical part Network), is the only ProtoPartNN that truly learns and localizes to prototypical object parts. We demonstrate that PixPNet achieves quantifiably improved interpretability without sacrificing accuracy.

****

LatentDR: Improving Model Generalization Through Sample-Aware Latent Degradation and Restoration

Ran Liu, Sahil Khose, Jingyun Xiao, Lakshmi Sathidevi, Keerthan Ramnath, Zsolt Kira, Eva L. Dyer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2669-2679

Despite significant advances in deep learning, models often struggle to generalize well to new, unseen domains, especially when training data is limited. To address this challenge, we propose a novel approach for distribution-aware latent augmentation that leverages the relationships across samples to guide the augmentation procedure. Our approach first degrades the samples stochastically in the latent space, mapping them to augmented labels, and then restores the samples from their corrupted versions during training. This process confuses the classifier in the degradation step and restores the overall class distribution of the original samples, promoting diverse intra-class/cross-domain variability. We extensively evaluate our approach on a diverse set of datasets and tasks, including domain generalization benchmarks and medical imaging datasets with strong domain shift, where we show our approach achieves significant improvements over existing methods for latent space augmentation. We further show that our method can be flexibly adapted to long-tail recognition tasks, demonstrating its versatility in building more generalizable models. Code is at https://github.com/nerdslab/LatentDR.

****

G-CASCADE: Efficient Cascaded Graph Convolutional Decoding for 2D Medical Image Segmentation

Md Mostafijur Rahman, Radu Marculescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7728-7737

In this paper, we are the first to propose a new graph convolution-based decoder namely, Cascaded Graph Convolutional Attention Decoder (G-CASCADE), for 2D medical image segmentation. G-CASCADE progressively refines multi-stage feature maps generated by hierarchical transformer encoders with an efficient graph convolution block. The encoder utilizes the self-attention mechanism to capture long-range dependencies, while the decoder refines the feature maps preserving long-range information due to the global receptive fields of the graph convolution block. Rigorous evaluations of our decoder with multiple transformer encoders on five medical image segmentation tasks (i.e., Abdomen organs, Cardiac organs, Polyp lesions, Skin lesions, and Retinal vessels) show that our model outperforms other state-of-the-art (SOTA) methods. We also demonstrate that our decoder achieves better DICE scores than the SOTA CASCADE decoder with 80.8% fewer parameters and 82.3% fewer FLOPs. Our decoder can easily be used with other hierarchical encoders for general-purpose semantic and medical image segmentation tasks. The implementation can be found at: https://github.com/SLDGroup/G-CASCADE.

****

TAMPAR: Visual Tampering Detection for Parcel Logistics in Postal Supply Chains

Alexander Naumann, Felix Hertlein, Laura Dörr, Kai Furmans; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8076-8086

Due to the steadily rising amount of valuable goods in supply chains, tampering detection for parcels is becoming increasingly important. In this work, we focus on the use-case last-mile delivery, where only a single RGB image is taken and compared against a reference from an existing database to detect potential appearance changes that indicate tampering. We propose a tampering detection pipeline that utilizes keypoint detection to identify the eight corner points of a parcel. This permits applying a perspective transformation to create normalized fronto-parallel views for each visible parcel side surface. These viewpoint-invariant parcel side surface representations facilitate the identification of signs of tampering on parcels within the supply chain, since they reduce the problem to parcel side surface matching with pair-wise appearance change detection. Experiments with multiple classical and deep learning-based change detection approaches are performed on our newly collected TAMpering detection dataset for PARcels, called TAMPAR. We evaluate keypoint and change detection separately, as well as in a unified system for tampering detection. Our evaluation shows promising results for keypoint (Keypoint AP 75.76) and tampering detection (81% accuracy, F1-Score 0.83) on real images. Furthermore, a sensitivity analysis for tampering types, lens distortion and viewing angles is presented. Code and dataset are available at https://a-nau.github.io/tampar.

****

PGVT: Pose-Guided Video Transformer for Fine-Grained Action Recognition

Haosong Zhang, Mei Chee Leong, Liyuan Li, Weisi Lin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6645-6656

Based on recent advancements in transformer-based video models and multi-modal joint learning, we propose a novel model, named Pose-Guided Video Transformer (PGVT), to incorporate sparse high-level body joints locations and dense low-level visual pixels for effective learning and accurate recognition of human actions. PGVT leverages the pre-trained image models by freezing their parameters and introducing trainable adapters to effectively integrate two input modalities, i.e., human poses and video frames, to learn a pose-focused spatiotemporal representation of human actions. We design two novel core modules, i.e., Pose Temporal Attention and Pose-Video Spatial Attention, to facilitate interaction between body joint locations and uniform video tokens, enriching each modality with contextualized information from the other. We evaluate PGVT model on four action recognition datasets: Diving48, Gym99, and Gym288 for fine-grained action recognition, and Kinetics400 for coarse-grained action recognition. Our model achieves new SOTA performance on the three fine-grained human action recognition datasets and comparable performance on Kinetics400 with a small number of tunable parameters compared with SOTA methods. The PGVT model exploits effective multi-modality learning by explicitly modeling human body joints and leveraging their contextualized interactions with video clips.

****

Multi-View Classification Using Hybrid Fusion and Mutual Distillation

Samuel Black, Richard Souvenir; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 270-280

Multi-view classification problems are common in medical image analysis, forensics, and other domains where problem queries involve multi-image input. Existing multi-view classification methods are often tailored to a specific task. In this paper, we repurpose off-the-shelf Hybrid CNN-Transformer networks for multi-view classification with either structured or unstructured views. Our approach incorporates a novel fusion scheme, mutual distillation, and introduces minimal additional parameters. We demonstrate the effectiveness and generalization capability of our approach, MV-HFMD, on multiple multi-view classification tasks and show that it outperforms other multi-view approaches, even task-specific methods. Code is available at https://github.com/vidarlab/multi-view-hybrid.

****

Real-Time User-Guided Adaptive Colorization With Vision Transformer

Gwanghan Lee, Saebyeol Shin, Taeyoung Na, Simon S. Woo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 484-493

Recently, the vision transformer (ViT) has achieved remarkable performance in computer vision tasks and has been actively utilized in colorization. Vision transformer uses multi-head self attention to effectively propagate user hints to distant relevant areas in the image. However, despite the success of vision transformers in colorizing the image, heavy underlying ViT architecture and the large computational cost hinder active real-time user interaction for colorization applications. Several research removed redundant image patches to reduce the computa tional cost of ViT in image classification tasks. However, the existing efficient ViT methods cause severe performance degradation in colorization task since it completely removes the redundant patches. Thus, we propose a novel efficient ViT architecture for real-time interactive colorization, AdaColViT determines which redundant image patches and layers to reduce in the ViT. Unlike existing methods, our novel pruning method alleviates performance drop and flexibly allocates computational resources of input samples, effectively achieving actual acceleration. In addition, we demonstrate through extensive experiments on ImageNet-ctest10k, Oxford 102flowers, and CUB-200 datasets that our method outperforms the baseline methods.

****

CAMOT: Camera Angle-Aware Multi-Object Tracking

Felix Limanta, Kuniaki Uto, Koichi Shinoda; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6479-6488

This paper proposes CAMOT, a simple camera angle estimator for multi-object tracking to tackle two problems: 1) occlusion and 2) inaccurate distance estimation in the depth direction. Under the assumption that multiple objects are located on a flat plane in each video frame, CAMOT estimates the camera angle using object detection. In addition, it gives the depth of each object, enabling pseudo-3D MOT. We evaluated its performance by adding it to various 2D MOT methods on the MOT17 and MOT20 datasets and confirmed its effectiveness. Applying CAMOT to ByteTrack, we obtained 63.8% HOTA, 80.6% MOTA, and 78.5% IDF1 in MOT17, which are state-of-the-art results. Its computational cost is significantly lower than the existing deep-learning-based depth estimators for tracking.

****

Egocentric Action Recognition by Capturing Hand-Object Contact and Object State

Tsukasa Shiota, Motohiro Takagi, Kaori Kumagai, Hitoshi Seshimo, Yushi Aono; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6541-6551

Improving the performance of egocentric action recognition (EAR) requires accurately capturing interactions between actors and objects. In this paper, we propose two learning methods that enable recognition models to capture hand object contact and object state change. We introduce Hand-Object Contact Learning (HOCL), which enables the model to focus on hand-object contact during actions, and Object State Learning (OSL), which enables the model to focus on object state changes caused by hand actions. Evaluation using a CNN-based model and a transformer-based model on the EGTEA, MECCANO, and EPIC-KITCHENS 100 datasets demonstrated the effectiveness of applying HOCL and OSL. Their application improved overall accuracy by up to 2.24% on EGTEA, 3.97% on MECCANO, and 1.49% on EPIC-KITCHENS 100. In addition, HOCL and OSL improved the performance on data with small training samples and one from unfamiliar scenes. Qualitative analysis revealed that their application enabled the models to precisely capture the interaction between actor and object.

****

IndustReal: A Dataset for Procedure Step Recognition Handling Execution Errors in Egocentric Videos in an Industrial-Like Setting

Tim J. Schoonbeek, Tim Houben, Hans Onvlee, Peter H.N. de With, Fons van der Sommen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4365-4374

Although action recognition for procedural tasks has received notable attention, it has a fundamental flaw in that no measure of success for actions is provided. This limits the applicability of such systems especially within the industrial domain, since the outcome of procedural actions is often significantly more important than the mere execution. To address this limitation, we define the novel task of procedure step recognition (PSR), focusing on recognizing the correct completion and order of procedural steps. Alongside the new task, we also present the multi-modal IndustReal dataset. Unlike currently available datasets, IndustReal contains procedural errors (such as omissions) as well as execution errors. A significant part of these errors are exclusively present in the validation and test sets, making IndustReal suitable to evaluate robustness of algorithms to n ew, unseen mistakes. Additionally, to encourage reproducibility and allow for scalable approaches trained on synthetic data, the 3D models of all parts are publicly available. Annotations and benchmark performance are provided for action recognition and assembly state detection, as well as the new PSR task. IndustReal, along with the code and model weights, is available at https://github.com/TimSchoonbeek/IndustReal.

****

FastCLIPstyler: Optimisation-Free Text-Based Image Style Transfer Using Style Representations

Ananda Padhmanabhan Suresh, Sanjana Jain, Pavit Noinongyao, Ankush Ganguly, Ukrit Watchareeruetai, Aubin Samacoits; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7316-7325

In recent years, language-driven artistic style transfer has emerged as a new type of style transfer technique, eliminating the need for a reference style image by using natural language descriptions of the style. The first model to achieve this, called CLIPstyler, has demonstrated impressive stylisation results. However, its lengthy optimisation procedure at runtime for each query limits its suitability for many practical applications. In this work, we present FastCLIPstyler, a generalised text-based image style transfer model capable of stylising images in a single forward pass for arbitrary text inputs. Furthermore, we introduce EdgeCLIPstyler, a lightweight model designed for compatibility with resource-constrained devices. Through quantitative and qualitative comparisons with state-of-the-art approaches, we demonstrate that our models achieve superior stylisation quality based on measurable metrics while offering significantly improved runtime efficiency, particularly on edge devices.

****

Video-kMaX: A Simple Unified Approach for Online and Near-Online Video Panoptic Segmentation

Inkyu Shin, Dahun Kim, Qihang Yu, Jun Xie, Hong-Seok Kim, Bradley Green, In So Kweon, Kuk-Jin Yoon, Liang-Chieh Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 229-239

Video Panoptic Segmentation (VPS) aims to achieve comprehensive pixel-level scene understanding by segmenting all pixels and associating objects in a video. Current solutions can be categorized into online and near-online approaches. Evolving over the time, each category has its own specialized designs, making it nontrivial to adapt models between different categories. To alleviate the discrepancy, in this work, we propose a unified approach for online and near-online VPS. The meta architecture of the proposed Video-kMaX consists of two components: within-clip segmenter (for clip-level segmentation) and cross-clip associater (for association beyond clips). We propose clip-kMaX (clip k-means mask transformer) and LA-MB (locationaware memory buffer) to instantiate the segmenter and associater, respectively. Our general formulation includes the online scenario as a special case by adopting clip length of one. Without bells and whistles, Video-kMaX sets a new state-of-the-art on KITTI-STEP and VIPSeg for video panoptic segmentation Code will be made publicly available. Code and models are available at this link: https://github.com/dlsrbgg33/video_kmax.

****

Cross-Feature Contrastive Loss for Decentralized Deep Learning on Heterogeneous Data

Sai Aparna Aketi, Kaushik Roy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 12-21

The current state-of-the-art decentralized learning algorithms mostly assume the data distribution to be Independent and Identically Distributed (IID). However, in practical scenarios, the distributed datasets can have significantly heterogeneous data distributions across the agents. In this work, we present a novel approach for decentralized learning on heterogeneous data, where data-free knowledge distillation through contrastive loss on cross-features is utilized to improve performance. Cross-features for a pair of neighboring agents are the features (i.e., last hidden layer activations) obtained from the data of an agent with respect to the model parameters of the other agent. We demonstrate the effectivene ss of the proposed technique through an exhaustive set of experiments on various Computer Vision datasets (CIFAR-10, CIFAR-100, Fashion MNIST, ImageNette, and ImageNet), model architectures, and network topologies. Our experiments show that the proposed method achieves superior performance (0.2-4% improvement in test accuracy) compared to other existing techniques for decentralized learning on heterogeneous data.

****

MOPA: Modular Object Navigation With PointGoal Agents

Sonia Raychaudhuri, Tommaso Campari, Unnat Jain, Manolis Savva, Angel X. Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5763-5773

We propose a simple but effective modular approach MOPA (Modular ObjectNav with PointGoal agents) to systematically investigate the inherent modularity of the object navigation task in Embodied AI. MOPA consists of four modules: (a) an object detection module trained to identify objects from RGB images, (b) a map building module to build a semantic map of the observed objects, (c) an exploration module enabling the agent to explore the environment, and (d) a navigation module to move to identified target objects. We show that we can effectively reuse a pretrained PointGoal agent as the navigation model instead of learning to navigate from scratch, thus saving time and compute. We also compare various exploration strategies for MOPA and find that a simple uniform strategy significantly outperforms more advanced exploration methods.

****

The Paleographer's Eye ex machina: Using Computer Vision To Assist Humanists in Scribal Hand Identification

Samuel Grieggs, C. E. M. Henderson, Sebastian Sobecki, Alexandra Gillespie, Walter Scheirer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7177-7186

The steady digitization of medieval manuscripts is rapidly changing the field of paleography, challenging existing assumptions about handwriting and book production. This development has identified historically important centers for the production of scribal texts, and even individual scribes themselves. For example, scholars of late medieval English literature have identified the copyists of a number of literary manuscripts, and the important role of London government clerks in shaping literary culture. However, traditional paleography has no agreed-upon methodology or fixed criteria for the attribution of handwriting to a particular community, period, or scribe. The approach taken by paleographers is inherently qualitative and subject to personal bias. Even those wielding the mighty "paleographer's eye" cannot claim objectivity. Computer vision offers solutions with spectacular performance on writer identification and retrieval benchmarks, but these have not been widely adopted by the paleography community because they tend not to hold up in practice. In this work, we attempt to bridge the divide with a software package designed not to automate paleography, but to augment the paleographer's eye. We introduce automated handwriting identification tools for which the results can be quickly visually understood and assessed, and used as one feature among many by expert paleographers when attributing previously unknown scribal hands. We also demonstrate a use case for our software by analyzing several items believed to be written by Thomas Hoccleve, a highly productive clerk of the Privy Seal who also happens to be an important fifteenth-century English poet.

****

Learning To Recognize Occluded and Small Objects With Partial Inputs

Hasib Zunair, A. Ben Hamza; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 675-684

Recognizing multiple objects in an image is challenging due to occlusions, and becomes even more so when the objects are small. While promising, existing multi-label image recognition models do not explicitly learn context-based representations, and hence struggle to correctly recognize small and occluded objects. Intuitively, recognizing occluded objects requires knowledge of partial input, and hence context. Motivated by this intuition, we propose Masked Supervised Learning

(MSL), a single-stage, model-agnostic learning paradigm for multi-label image recognition. The key idea is to learn context-based representations using a masked branch and to model label co-occurrence using label consistency. Experimental results demonstrate the simplicity, applicability and more importantly the competitive performance of MSL against previous state-of-the-art methods on standard multi-label image recognition benchmarks. In addition, we show that MSL is robust to random masking and demonstrate its effectiveness in recognizing non-masked objects. Code and pretrained models are available on GitHub.

****

BALF: Simple and Efficient Blur Aware Local Feature Detector

Zhenjun Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3362-3372

Local feature detection is a key ingredient of many image processing and computer vision applications, such as visual odometry and localization. Most existing algorithms focus on feature detection from a sharp image. They would thus have degraded performance once the image is blurred, which could happen easily under low-lighting conditions. To address this issue, we propose a simple yet both efficient and effective keypoint detection method that is able to accurately localize the salient keypoints in a blurred image. Our method takes advantages of a novel multi-layer perceptron (MLP) based architecture that significantly improve the detection repeatability for a blurred image. The network is also light-weight and able to run in real-time, which enables its deployment for time-constrained applications. Extensive experimental results demonstrate that our detector is able to improve the detection repeatability with blurred images, while keeping comparable performance as existing state-of-the-art detectors for sharp images. The code and trained weights are publicly available at github.com/ericzzj1989/BALF.

****

RS2G: Data-Driven Scene-Graph Extraction and Embedding for Robust Autonomous Perception and Scenario Understanding

Junyao Wang, Arnav Vaibhav Malawade, Junhong Zhou, Shih-Yuan Yu, Mohammad Abdullah Al Faruque; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7493-7502

Effectively capturing intricate interactions among road users is of critical importance to achieving safe navigation for autonomous vehicles. While graph learning (GL) has emerged as a promising approach to tackle this challenge, existing GL models rely on predefined domain-specific graph extraction rules that often fail in real-world drastically changing scenarios. Additionally, these graph extraction rules severely impede the capability of existing GL methods to generalize knowledge across domains. To address this issue, we propose RoadScene2Graph (RS2G), an innovative autonomous scenario understanding framework with a novel data-driven graph extraction and modeling approach that dynamically captures the diverse relations among road users. Our evaluations demonstrate that on average RS2G outperforms the state-of-the-art (SOTA) rule-based graph extraction method by 4.47% and the SOTA deep learning model by 22.19% in subjective risk assessment. More importantly, RS2G delivers notably better performance in transferring knowledge gained from simulation environments to unseen real-world scenarios.

****

Leveraging the Power of Data Augmentation for Transformer-Based Tracking

Jie Zhao, Johan Edstedt, Michael Felsberg, Dong Wang, Huchuan Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6469-6478

Due to long-distance correlation and powerful pretrained models, transformer-based methods have initiated a breakthrough in visual object tracking performance. Previous works focus on designing effective architectures suited for tracking, but ignore that data augmentation is equally crucial for training a well-performing model. In this paper, we first explore the impact of general data augmentations on transformer-based trackers via systematic experiments, and reveal the limited effectiveness of these common strategies. Motivated by experimental observations, we then propose two data augmentation methods customized for tracking. First, we optimize existing random cropping via a dynamic search radius mechanism a nd simulation for boundary samples. Second, we propose a token-level feature mixing augmentation strategy, which enables the model against challenges like background interference. Extensive experiments on two transformer-based trackers and six benchmarks demonstrate the effectiveness and data efficiency of our methods, especially under challenging settings, like one-shot tracking and small image resolutions. Code is available at https://github.com/zj5559/DATr.

****

Med-DANet V2: A Flexible Dynamic Architecture for Efficient Medical Volumetric Segmentation

Haoran Shen, Yifu Zhang, Wenxuan Wang, Chen Chen, Jing Liu, Shanshan Song, Jiangyun Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7871-7881

Recent works have shown that the computational efficiency of 3D medical image (e.g. CT and MRI) segmentation can be impressively improved by dynamic inference based on slice-wise complexity. As a pioneering work, a dynamic architecture network for medical volumetric segmentation (i.e. Med-DANet) has achieved a favorable accuracy and efficiency trade-off by dynamically selecting a suitable 2D candidate model from the pre-defined model bank for different slices. However, the issues of incomplete data analysis, high training costs, and the two-stage pipeline in Med-DANet require further improvement. To this end, this paper further explores a unified formulation of the dynamic inference framework from the perspective of both the data itself and the model structure. For each slice of the input volume, our proposed method dynamically selects an important foreground region for segmentation based on the policy generated by our Decision Network and Crop Position Network. Besides, we propose to insert a stage-wise quantization selector to the employed segmentation model (e.g. U-Net) for dynamic architecture adapting. Extensive experiments on BraTS 2019 and 2020 show that our method achieves comparable or better performance than previous state-of-the-art methods with much less model complexity. Compared with previous methods Med-DANet and TransBTS with dynamic and static architecture respectively, our framework improves the model efficiency by up to nearly 4.1 and 17.3 times with comparable segmentation results on BraTS 2019. Code will be available at https://github.com/Rubics-Xuan/Med-DANet.

****

Partial Binarization of Neural Networks for Budget-Aware Efficient Learning

Udbhav Bamba, Neeraj Anand, Saksham Aggarwal, Dilip K. Prasad, Deepak K. Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2336-2345

Binarization is a powerful compression technique for neural networks, significantly reducing FLOPs, but often results in a significant drop in model performance. To address this issue, partial binarization techniques have been developed, but a systematic approach to mixing binary and full-precision parameters in a single network is still lacking. In this paper, we propose a controlled approach to partial binarization, creating a budgeted binary neural network (B2NN) with our MixBin strategy. This method optimizes the mixing of binary and full-precision components, allowing for explicit selection of the fraction of the network to remain binary. Our experiments show that B2NNs created using MixBin outperform those from random or iterative searches and state-of-the-art layer selection methods by up to 3% on the ImageNet-1K dataset. We also show that B2NNs outperform the structured pruning baseline by approximately 23% at the extreme FLOP budget of 15%, and perform well in object tracking, with up to a 12.4% relative improvement over other baselines. Additionally, we demonstrate that B2NNs developed by MixBin can be transferred across datasets, with some cases showing improved performance over directly applying MixBin on the downstream data.

****

Improving the Fairness of the Min-Max Game in GANs Training

Zhaoyu Zhang, Yang Hua, Hui Wang, Seán McLoone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2910-2919

Generative adversarial networks (GANs) have achieved great success and become more and more popular in recent years. However, understanding of the min-max game in GANs training is still limited. In this paper, we first utilize information game theory to analyze the min-max game in GANs and introduce a new viewpoint on the GANs training that the min-max game in existing GANs is unfair during training, leading to sub-optimal convergence. To tackle this, we propose a novel GAN called Information Gap GAN (IGGAN), which consists of one generator (G) and two discriminators (D1 and D2). Specifically, we apply different data augmentation methods to D1 and D2, respectively. The information gap between different data augmentation methods can change the information received by each player in the min-max game and lead to all three players G, D1 and D2 in IGGAN obtaining incomplete information, which improves the fairness of the min-max game, yielding better convergence. We conduct extensive experiments for large-scale and limited data settings on several common datasets with two backbones, i.e., BigGAN and StyleGAN2. The results demonstrate that IGGAN can achieve a higher Inception Score (IS) and a lower Frechet Inception Distance (FID) compared with other GANs. Codes are available at https://github.com/zzhang05/IGGAN

****

When 3D Bounding-Box Meets SAM: Point Cloud Instance Segmentation With Weak-and-Noisy Supervision

Qingtao Yu, Heming Du, Chen Liu, Xin Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3719-3728

Learning from bounding-boxes annotations has shown great potential in weakly-supervised 3D point cloud in- stance segmentation. However, we observed that existing methods would suffer severe performance degradation with perturbed bounding box annotations. To tackle this is- sue, we propose a complementary image prompt-induced weakly-supervised point cloud instance segmentation (CIP- WPIS) method. CIP-WPIS leverages pretrained knowledge embedded in the 2D foundation model SAM and 3D geo- metric prior to achieve accurate point-wise instance labels from the bounding box annotations. Specifically, CIP-WPIS first selects image views in which 3D candidate points of an instance are fully visible. Then, we generate complemen- tary background and foreground prompts from projections to obtain SAM 2D instance mask predictions. According to these, we assign the confidence values to points indicating the likelihood of points belonging to the instance. Furthermore, we utilize 3D geometric homogeneity provided by superpoints to decide the final instance label assignments. In this fashion, we achieve high-quality 3D point-wise in- stance labels. Extensive experiments on both Scannet-v2 and S3DIS benchmarks proves that our method not only achieves state-of-the-art performance for bounding-boxes supervised point cloud instance segmentation, but also exhibits robustness against noisy 3D bounding-box annotations.

****

Domain Aligned CLIP for Few-Shot Classification

Muhammad Waleed Gondal, Jochen Gast, Inigo Alonso Ruiz, Richard Droste, Tommaso Macri, Suren Kumar, Luitpold Staudigl; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5721-5730

Large vision-language representation learning models like CLIP have demonstrated impressive performance for zero-shot transfer to downstream tasks while largely benefiting from inter-modal (image-text) alignment via contrastive objectives. This downstream performance can further be enhanced by full-scale fine-tuning which is often compute intensive, requires large labelled data, and can reduce out-of-distribution (OOD) robustness. Furthermore, sole reliance on inter-modal alignment might overlook the rich information embedded within each individual modality. In this work, we introduce a sample-efficient domain adaptation strategy for CLIP, termed Domain Aligned CLIP (DAC), which improves both intra-modal (image-image) and inter-modal alignment on target distributions without fine-tuning the main model. For intra-modal alignment, we introduce a lightweight adapter that is specifically trained with an intra-modal contrastive objective. To improve inter-modal alignment, we introduce a simple framework to modulate the precomputed class text embeddings. The proposed few-shot fine-tuning framework is computationally efficient, robust to distribution shifts, and does not alter CLIP's parameters. We study the effectiveness of DAC by benchmarking on 11 widely used image classification tasks with consistent improvements in 16-shot classification up on strong baselines by about 2.3% and demonstrate competitive performance on 4 OOD robustness benchmarks.

****

Beyond Document Page Classification: Design, Datasets, and Challenges

Jordy Van Landeghem, Sanket Biswas, Matthew Blaschko, Marie-Francine Moens; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2962-2972

This paper highlights the need to bring document classification benchmarking closer to real-world applications, both in the nature of data tested (X: multi-channel, multi-paged, multi-industry; Y: class distributions and label set variety) and in classification tasks considered (f: multi-page document, page stream, and document bundle classification, ...). We identify the lack of public multi-page document classification datasets, formalize different classification tasks arising in application scenarios, and motivate the value of targeting efficient multi-page document representations. An experimental study on proposed multi-page document classification datasets demonstrates that current benchmarks have become irrelevant and need to be updated to evaluate complete documents, as they naturally occur in practice. This reality check also calls for more mature evaluation methodologies, covering calibration evaluation, inference complexity (time-memory), and a range of realistic distribution shifts (e.g., born-digital vs. scanning noise, shifting page order). Our study ends on a hopeful note by recommending concrete avenues for future improvements.

****

Towards More Realistic Membership Inference Attacks on Large Diffusion Models

Jan Dubi■ski, Antoni Kowalczuk, Stanis■aw Pawlak, Przemyslaw Rokita, Tomasz Trzci■ski, Pawe■ Morawiecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4860-4869

Generative diffusion models, including Stable Diffusion and Midjourney, can generate visually appealing, diverse, and high-resolution images for various applications. These models are trained on billions of internet-sourced images, raising significant concerns about the potential unauthorized use of copyright-protected images. In this paper, we examine whether it is possible to determine if a specific image was used in the training set, a problem known as a membership inference attack. Our focus is on Stable Diffusion, and we address the challenge of designing a fair evaluation framework to answer this membership question. We propose a new dataset to establish a fair evaluation setup and apply it to Stable Diffusion, also applicable to other generative models. With the proposed dataset, we execute membership attacks (both known and newly introduced). Our research reveals that previously proposed evaluation setups do not provide a full understanding of the effectiveness of membership inference attacks. We conclude that the membership inference attack remains a significant challenge for large diffusion models (often deployed as black-box systems), indicating that related privacy and copyright issues will persist in the foreseeable future.

****

Slice and Conquer: A Planar-to-3D Framework for Efficient Interactive Segmentation of Volumetric Images

Wonwoo Cho, Dongmin Choi, Hyesu Lim, Jinho Choi, Saemee Choi, Hyun-seok Min, Sungbin Lim, Jaegul Choo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7614-7623

Interactive segmentation methods have been investigated to address the potential need for additional refinement in automatic segmentation via human-in-the-loop techniques. For accurate segmentation of 3D images, we propose Slice-and-Conquer, a novel planar-to-3D pipeline formulating volumetric mask construction into two stages: 1) 2D interactive segmentation and 2) guided 3D segmentation. Specifically, the first stage enables users to focus on a single 2D slice and provides the corresponding 2D prediction results as strong shape priors. Taking the planar guidance, an accurate 3D mask can be constructed with minimal interactions. To support a flexible iterative refinement, our system recommends a next slice to annotate at the end of the second stage. Since volumetric segmentation can be completed by consecutively annotating a few recommended 2D slices, our method signi ficantly reduces the cognitive burden of exploring volumetric space for users. Through extensive experiments on various datasets of 3D biomedical images, we demonstrate the effectiveness of the proposed pipeline.

****

Mitigate Domain Shift by Primary-Auxiliary Objectives Association for Generalizing Person ReID

Qilei Li, Shaogang Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 394-403

While deep learning has significantly improved ReID model accuracy under the independent and identical distribution (IID) assumption, it has also become clear that such models degrade notably when applied to an unseen novel domain due to unpredictable/unknown domain shift. Contemporary domain generalization (DG) ReID models struggle in learning domain-invariant representation through solely training on an instance classification objective. We consider that a deep learning model is heavily influenced therefore biased towards domain-specific characteristics, e.g., background clutter, scale and viewpoint variations, limiting the generalizability of the learned model, and hypothesize that the pedestrians are domain invariant owning they share the same structural characteristics. To enable ReID model to be less domain-specific from these pure pedestrians and domain-specific factors, we introduce a method that guides model learning of the primary ReID instance classification objective by a concurrent auxiliary learning objective on weakly labeled pedestrian saliency detection. To solve the problem of conflicting optimization criteria in the model parameter space between the two learning objectives, we introduce a Primary-Auxiliary Objectives Association (PAOA) mechanism to calibrate the loss gradients of the auxiliary task towards the primary learning task gradients. Benefited from the harmonious multitask learning design, our model can be extended with the recent test-time diagram to form the PAOA+, which performs on-the-fly optimization against the auxiliary objective in order to maximize the model's generative capacity in the test target domain. Experiments demonstrate the superiority of the proposed PAOA model.

****

MonoProb: Self-Supervised Monocular Depth Estimation With Interpretable Uncertainty

Rémi Marsal, Florian Chabot, Angélique Loesch, William Grolleau, Hichem Sahbi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3637-3646

Self-supervised monocular depth estimation methods aim to be used in critical applications such as autonomous vehicles for environment analysis. To circumvent the potential imperfections of these approaches, a quantification of the prediction confidence is crucial to guide decision-making systems that rely on depth estimation. In this paper, we propose MonoProb, a new unsupervised monocular depth estimation method that returns an interpretable uncertainty, which means that the uncertainty reflects the expected error of the network in its depth predictions. We rethink the stereo or the structure-from-motion paradigms used to train unsupervised monocular depth models as a probabilistic problem. Within a single forward pass inference, this model provides a depth prediction and a measure of its confidence, without increasing the inference time. We then improve the performance on depth and uncertainty with a novel self-distillation loss for which a student is supervised by a pseudo ground truth that is a probability distribution on depth output by a teacher. To quantify the performance of our models we design new metrics that, unlike traditional ones, measure the absolute performance of uncertainty predictions. Our experiments highlight enhancements achieved by our method on standard depth and uncertainty metrics as well as on our tailored metrics. https://github.com/CEA-LIST/MonoProb

****

LP-OVOD: Open-Vocabulary Object Detection by Linear Probing

Chau Pham, Truong Vu, Khoi Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 779-788

This paper addresses the challenging problem of open-vocabulary object detection (OVOD) where an object detector must identify both seen and unseen classes in t est images without labeled examples of the unseen classes in training. A typical approach for OVOD is to use joint text-image embeddings of CLIP to assign box proposals to their closest text label. However, this method has a critical issue: many low-quality boxes, such as over- and under-covered-object boxes, have the same similarity score as high-quality boxes since CLIP is not trained on exact object location information. To address this issue, we propose a novel method, LP-OVOD, that discards low-quality boxes by training a sigmoid linear classifier on pseudo labels retrieved from the top relevant region proposals to the novel text. Notably, LP-OVOD seamlessly integrates the knowledge distillation technique from ViLD, resulting in a new state-of-the-art OVOD approach. Experimental results on COCO affirm the superior performance of our approach over prior work, achieving 40.5 in AP_novel using ResNet50 as the backbone and without external datasets or knowing novel classes in training. Our code will be available at https://github.com/VinAIResearch/LP-OVOD.

****

Beyond Active Learning: Leveraging the Full Potential of Human Interaction via Auto-Labeling, Human Correction, and Human Verification

Nathan Beck, Krishnateja Killamsetty, Suraj Kothawade, Rishabh Iyer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2881-2889

Active Learning (AL) is a human-in-the-loop framework to interactively and adaptively label data instances, thereby enabling significant gains in model performance compared to random sampling. AL approaches function by selecting the hardest instances to label, often relying on notions of diversity and uncertainty. However, we believe that these current paradigms of AL do not leverage the full potential of human interaction granted by automated label suggestions. Indeed, we show that for many classification tasks and datasets, most people verifying if an automatically suggested label is correct take 3x to 4x less time than they do changing an incorrect suggestion to the correct label (or labeling from scratch without any suggestion). Utilizing this result, we propose CLARIFIER (aCtive LeARnIng From tIEred haRdness), an Interactive Learning framework that admits more effective use of human interaction by leveraging the reduced cost of verification. By targeting the hard (uncertain) instances with existing AL methods, the intermediate instances with a novel label suggestion scheme using submodular mutual information functions on a per-class basis, and the easy (confident) instances with highest-confidence auto-labeling, CLARIFIER can improve over the performance of existing AL approaches on multiple datasets -- particularly on those that have a large number of classes -- by almost 1.5x to 2x in terms of relative labeling cost.

****

ARNIQA: Learning Distortion Manifold for Image Quality Assessment

Lorenzo Agnolucci, Leonardo Galteri, Marco Bertini, Alberto Del Bimbo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 189-198

No-Reference Image Quality Assessment (NR-IQA) aims to develop methods to measure image quality in alignment with human perception without the need for a high-quality reference image. In this work, we propose a self-supervised approach named ARNIQA (leArning distoRtion maNifold for Image Quality Assessment) for modeling the image distortion manifold to obtain quality representations in an intrinsic manner. First, we introduce an image degradation model that randomly composes ordered sequences of consecutively applied distortions. In this way, we can synthetically degrade images with a large variety of degradation patterns. Second, we propose to train our model by maximizing the similarity between the representations of patches of different images distorted equally, despite varying content. Thus, images degraded in the same manner correspond to neighboring positions within the distortion manifold. Finally, we map the image representations to the quality scores with a simple linear regressor, thus without fine-tuning the encoder weights. The experiments show that our approach achieves state-of-the-art performance on several datasets. In addition, ARNIQA demonstrates improved data efficiency, generalization capabilities, and robustness compared to competing metho ds. The code and the model are publicly available at https://github.com/miccunifi/ARNIQA.

****

CVTHead: One-Shot Controllable Head Avatar With Vertex-Feature Transformer

Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6131-6141

Reconstructing personalized animatable head avatars has significant implications in the fields of AR/VR. Existing methods for achieving explicit face control of 3D Morphable Models (3DMM) typically rely on multi-view images or videos of a single subject, making the reconstruction process complex. Additionally, the traditional rendering pipeline is time-consuming, limiting real-time animation possibilities. In this paper, we introduce CVTHead, a novel approach that generates controllable neural head avatars from a single reference image using point-based neural rendering. CVTHead considers the sparse vertices of mesh as the point set and employs the proposed Vertex-feature Transformer to learn local feature descriptors for each vertex. This enables the modeling of long-range dependencies among all the vertices. Experimental results on the VoxCeleb dataset demonstrate that CVTHead achieves comparable performance to state-of-the-art graphics-based methods. Moreover, it enables efficient rendering of novel human heads with various expressions, head poses, and camera views. These attributes can be explicitly controlled using the coefficients of 3DMMs, facilitating versatile and realistic animation in real-time scenarios.

****

FIRe: Fast Inverse Rendering Using Directional and Signed Distance Functions

Tarun Yenamandra, Ayush Tewari, Nan Yang, Florian Bernard, Christian Theobalt, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3077-3087

Neural 3D implicit representations learn priors that are useful for diverse applications, such as single- or multiple-view 3D reconstruction. A major downside of existing approaches while rendering an image is that they require evaluating the network multiple times per camera ray so that the high computational time forms a bottleneck for downstream applications. We address this problem by introducing a novel neural scene representation that we call the directional distance function (DDF). To this end, we learn a signed distance function (SDF) along with our DDF model to represent a class of shapes. Specifically, our DDF is defined on the unit sphere and predicts the distance to the surface along any given direction. Therefore, our DDF allows rendering images with just a single network evaluation per camera ray. Based on our DDF, we present a novel fast algorithm (FIRe) to reconstruct 3D shapes given a posed depth map. We evaluate our proposed method on 3D reconstruction from single-view depth images, where we empirically show that our algorithm reconstructs 3D shapes more accurately and it is more than 15 times faster (per iteration) than competing methods.

****

Ego2HandsPose: A Dataset for Egocentric Two-Hand 3D Global Pose Estimation

Fanqing Lin, Tony Martinez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4375-4383

Color-based two-hand 3D pose estimation in the global coordinate system is essential in many applications. However, there are very few datasets dedicated to this task and no existing dataset supports estimation in a non-laboratory environment. This is largely attributed to the sophisticated data collection process required for 3D hand pose annotations, which also leads to difficulty in obtaining instances with the level of visual diversity needed for estimation in the wild. Progressing towards this goal, a large-scale dataset Ego2Hands was recently proposed to address the task of two-hand segmentation and detection in the wild. The proposed composition-based data generation technique can create two-hand instances with quality, quantity and diversity that generalize well to unseen domains. In this work, we present Ego2HandsPose, an extension of Ego2Hands that contains 3D hand pose annotation and is the first dataset that enables color-based two-hand 3D tracking in unseen domains. To this end, we develop a set of parametric fi tting algorithms to enable 1) 3D hand pose annotation using a single image, 2) automatic conversion from 2D to 3D hand poses and 3) accurate two-hand tracking with temporal consistency. We provide incremental quantitative analysis on the multi-stage pipeline and show that training on our dataset achieves state-of-the-art results that significantly outperforms other datasets for the task of egocentric two-hand global 3D pose estimation.

****

Improving Vision-and-Language Reasoning via Spatial Relations Modeling

Cheng Yang, Rui Xu, Ye Guo, Peixiang Huang, Yiru Chen, Wenkui Ding, Zhongyuan Wang, Hong Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 769-778

Visual commonsense reasoning (VCR) is a challenging multi-modal task, which requires high-level cognition and commonsense reasoning ability about the real world. In recent years, large-scale pre-training approaches have been developed and promoted the state-of-the-art performance of VCR. However, the existing approaches almost employ the BERT-like objectives to learn multi-modal representations. These objectives motivated from the text-domain are insufficient for the excavation on the complex scenario of visual modality. Most importantly, the spatial distribution of the visual objects is basically neglected. To address the above issue, we propose to construct the spatial relation graph based on the given visual scenario. Further, we design two pre- training tasks named object position regression (OPR) and spatial relation classification (SRC) to learn to reconstruct the spatial relation graph respectively. Quantitative analysis suggests that the proposed method can guide the representations to maintain more spatial context and facilitate the attention on the essential visual regions for reasoning. We achieve the state-of-the-art results on VCR and two other vision-and-language reasoning tasks VQA, and NLVR2.

****

WATCH: Wide-Area Terrestrial Change Hypercube

Connor Greenwell, Jon Crall, Matthew Purri, Kristin Dana, Nathan Jacobs, Armin Hadzic, Scott Workman, Matt Leotta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8277-8286

Monitoring Earth activity using data collected from multiple satellite imaging platforms in a unified way is a significant challenge, especially with large variability in image resolution, spectral bands, and revisit rates. Further, the availability of sensor data varies across time as new platforms are launched. In this work, we introduce an adaptable framework and network architecture capable of predicting on subsets of the available platforms, bands, or temporal ranges it was trained on. Our system, called WATCH, is highly general and can be applied to a variety of geospatial tasks. In this work, we analyze the performance of WATCH using the recent IARPA SMART public dataset and metrics. We focus primarily on the problem of broad area search for heavy construction sites. Experiments validate the robustness of WATCH during inference to limited sensor availability, as well the the ability to alter inference-time spatial or temporal sampling. WATCH is open source and available for use on this or other remote sensing problems. Code and model weights are available at: https://gitlab.kitware.com/computer-vision/geowatch

****

Detecting Content Segments From Online Sports Streaming Events: Challenges and Solutions

Zongyi Liu, Yarong Feng, Shunyan Luo, Yuan Ling, Shujing Dong, Shuyi Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6414-6425

Developing a client-side segmentation algorithm for online sports streaming holds significant importance. For instance, in order to assess the video quality from an end-user perspective such as artifact detection, it is important to initially segment the content within the streaming playback. The challenge lies in localizing the content due to the intricate scene changes between content and non-content sections in popular sports like football, tennis, baseball, and more. Client-side content detection can be implemented in two ways: intrusively, involving the interception of network traffic and parsing service provider data and logs, or non-intrusively, which entails capturing streamed videos from content providers and subjecting them to analysis using computer vision technologies. In this paper, we introduce a non-intrusive framework that leverages a combination of traditional machine learning algorithms and deep neural networks (DNN) to distinguish content sections from non-content sections across various online sports streaming services. Our algorithm has demonstrated a remarkable level of accuracy and effectiveness in sports broadcasting events, effectively overcoming the complexities introduced by intricate non-content insertion methods during the games.

****

Vikriti-ID: A Novel Approach for Real Looking Fingerprint Data-Set Generation

Rishabh Shukla, Aditya Sinha, Vansh Singh, Harkeerat Kaur; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6395-6403

Fingerprint recognition research faces significant challenges due to the limited availability of extensive and publicly available fingerprint databases. Existing databases lack a sufficient number of identities and fingerprint impressions, which hinders progress in areas such as Fingerprintbased access control. To address this challenge, we present Vikriti-ID, a synthetic fingerprint generator capable of generating unique fingerprints with multiple impressions. Using Vikriti-ID, we generated a large database containing 500000 unique fingerprints, each with 10 associated impressions. We then demonstrate the effectiveness of the database generated by Vikriti-ID by evaluating it for imposter-genuine score distribution and EER score. Apart from this we also trained a deep network to check the usability of data. We trained a deep network on both Vikriti-ID generated data as well as public data. This generated data achieved an Equal Error Rate(EER) of 0.16%, AUC of 0.89%. This improvement is possible due to the limitations of existing publicly available data-set, which struggle in numbers or multiple impressions.

****

PETIT-GAN: Physically Enhanced Thermal Image-Translating Generative Adversarial Network

Omri Berman, Navot Oz, David Mendlovic, Nir Sochen, Yafit Cohen, Iftach Klapp; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1618-1627

Thermal multispectral imagery is imperative for a plethora of environmental applications. Unfortunately, there are no publicly-available datasets of thermal multispectral images with a high spatial resolution that would enable the development of algorithms and systems in this field. However, image-to-image (I2I) translation could be used to artificially synthesize such data by transforming largely-available datasets of other visual modalities. In most cases, pairs of content-wise-aligned input-target images are not available, making it harder to train and converge to a satisfying solution. Nevertheless, some data domains, and particularly the thermal domain, have unique properties that tie the input to the output that could help mitigate those weaknesses. We propose PETIT-GAN, a physically enhanced thermal image-translating generative adversarial network to transform between different thermal modalities - a step toward synthesizing a complete thermal multispectral dataset. Our novel approach embeds physically modeled prior information in an UI2I translation to produce outputs with greater fidelity to the target modality. We further show that our solution outperforms the current state-of-the-art architectures at thermal UI2I translation by approximately 50% with respect to the standard perceptual metrics, and enjoys a more robust training procedure. The code and data used for the development and analysis of our method are publicly available and can be accessed through our project's website: https://bermanz.github.io/PETIT

****

Design Choices for Enhancing Noisy Student Self-Training

Aswathnarayan Radhakrishnan, Jim Davis, Zachary Rabin, Benjamin Lewis, Matthew Scherreik, Roman Ilin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1926-1935

Semi-supervised learning approaches train on small sets of labeled data in addition to large sets of unlabeled data. Self-training is a semi-supervised teacher-student approach that often suffers from "confirmation bias" that occurs when the student model repeatedly overfits to incorrect pseudo-labels given by the teacher model for the unlabeled data. This bias impedes improvements in pseudo-label accuracy across self-training iterations, leading to unwanted saturation in model performance after just a few iterations. In this work, we study multiple design choices to improve the Noisy Student self-training pipeline and reduce confirmation bias. We showed that our proposed Weighted SplitBatch Sampler and Dataset-Adaptive Techniques for Model Calibration and Entropy-Based Pseudo-Label Selection provided performance gains over existing design choices across multiple datasets. Finally, we also study the extendability of our enhanced approach to Open Set unlabeled data (containing classes not seen in labeled data). The source code can be licensed for use via email.

****

ArcGeo: Localizing Limited Field-of-View Images Using Cross-View Matching

Maxim Shugaev, Ilya Semenov, Kyle Ashley, Michael Klaczynski, Naresh Cuntoor, Mun Wai Lee, Nathan Jacobs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 209-218

Cross-view matching techniques for image geolocalization attempt to match features in ground level imagery against a collection of satellite images to determine the position of given query image. We present a novel cross-view image matching approach called ArcGeo which introduces a batch-all angular margin loss and several train-time strategies including large-scale pretraining and FoV-based data augmentation. This allows our model to perform well even in challenging cases with limited field-of-view (FoV). Further, we evaluate multiple model architectures, data augmentation approaches and optimization strategies to train a deep cross-view matching network, specifically optimized for limited FoV cases. In low FoV experiments (FoV = 90deg) our method improves top-1 image recall rate on the CVUSA dataset from 30.12% to 43.08%. We also demonstrate improved performance over the state-of-the-art techniques for panoramic cross-view retrieval, improving top-1 recall from 95.43% to 96.06% on the CVUSA dataset and from 64.52% to 79.88% on the CVACT test dataset. Lastly, we evaluate the role of large-scale pretraining for improved robustness. With appropriate pretraining on external data, our model improves top-1 recall dramatically to 66.83% for FoV = 90deg test case on CVUSA, an increase of over twice what is reported by existing approaches.

****

Understanding Hyperbolic Metric Learning Through Hard Negative Sampling

Yun Yue, Fangzhou Lin, Guanyi Mou, Ziming Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1891-1903

In recent years, there has been a growing trend of incorporating hyperbolic geometry methods into computer vision. While these methods have achieved state-of-the-art performance on various metric learning tasks using hyperbolic distance measurements, the underlying theoretical analysis supporting this superior performance remains under-exploited. In this study, we investigate the effects of integrating hyperbolic space into metric learning, particularly when training with contrastive loss. We identify a need for a comprehensive comparison between Euclidean and hyperbolic spaces regarding the temperature effect in the contrastive loss within the existing literature. To address this gap, we conduct an extensive investigation to benchmark the results of Vision Transformers (ViTs) using a hybrid objective function that combines loss from Euclidean and hyperbolic spaces. Additionally, we provide a theoretical analysis of the observed performance improvement. We also reveal that hyperbolic metric learning is highly related to hard negative sampling, providing insights for future work. This work will provide valuable data points and experience in understanding hyperbolic image embeddings. To shed more light on problem-solving and encourage further investigation into our approach, our code is available online.

****

FIRE: Food Image to REcipe Generation

Prateek Chhikara, Dhiraj Chaurasia, Yifan Jiang, Omkar Masur, Filip Ilievski; Pr oceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8184-8194

Food computing has emerged as a prominent multidisciplinary field of research in recent years. An ambitious goal of food computing is to develop end-to-end intelligent systems capable of autonomously producing recipe information for a food image. Current image-to-recipe methods are retrieval-based and their success depends heavily on the dataset size and diversity, as well as the quality of learned embeddings. Meanwhile, the emergence of powerful attention-based vision and language models presents a promising avenue for accurate and generalizable recipe generation, which has yet to be extensively explored. This paper proposes FIRE, a novel multimodal methodology tailored to recipe generation in the food computing domain, which generates the food title, ingredients, and cooking instructions based on input food images. FIRE leverages the BLIP model to generate titles, utilizes a Vision Transformer with a decoder for ingredient extraction, and employs the T5 model to generate recipes incorporating titles and ingredients as inputs. We showcase two practical applications that can benefit from integrating FIRE with large language model prompting: recipe customization to fit recipes to user preferences and recipe-to-code transformation to enable automated cooking processes. Our experimental findings validate the efficacy of our proposed approach, underscoring its potential for future advancements and widespread adoption in food computing.

****

DiffCLIP: Leveraging Stable Diffusion for Language Grounded 3D Classification

Sitian Shen, Zilin Zhu, Linqian Fan, Harry Zhang, Xinxiao Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3596-3605

Large pre-trained models have revolutionized the field of computer vision by facilitating multi-modal learning. Notably, the CLIP model has exhibited remarkable proficiency in tasks such as image classification, object detection, and semantic segmentation. Nevertheless, its efficacy in processing 3D point clouds is restricted by the domain gap between the depth maps derived from 3D projection and the training images of CLIP. This paper introduces DiffCLIP, a novel pre-training framework that seamlessly integrates stable diffusion with ControlNet. The primary objective of DiffCLIP is to bridge the domain gap inherent in the visual branch. Furthermore, to address few-shot tasks in the textual branch, we incorporate a style-prompt generation module. Extensive experiments on the ModelNet10, ModelNet40, and ScanObjectNN datasets show that DiffCLIP has strong abilities for 3D understanding. By using stable diffusion and style-prompt generation, DiffCLIP achieves an accuracy of 43.2% for zero-shot classification on OBJ_BG of ScanObjectNN, which is state-of-the-art performance, and an accuracy of 82.4% for zero-shot classification on ModelNet10, which is also state-of-the-art performance.

****

A One-Shot Learning Approach To Document Layout Segmentation of Ancient Arabic Manuscripts

Axel De Nardin, Silvia Zottin, Claudio Piciarelli, Emanuela Colombi, Gian Luca Foresti; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8127-8136

Document layout segmentation is a challenging task due to the variability and complexity of document layouts. Ancient manuscripts in particular are often damaged by age, have very irregular layouts, and are characterized by progressive editing from different authors over a large time window. All these factors make the semantic segmentation process of specific areas, such as main text and side text, very difficult. However, the study of these manuscripts turns out to be fundamental for historians and humanists, so much so that in recent years the demand for machine learning approaches aimed at simplifying the extraction of information from these documents has consistently increased, leading document layout analysis to become an increasingly important research area. In order for machine learning techniques to be applied effectively to this task, however, a large amount of correctly and precisely labeled images is required for their training. This is obviously a limitation for this field of research as ground truth must be prec isely and manually crafted by expert humanists, making it a very time-consuming process. In this paper, with the aim of overcoming this limitation, we present an efficient document layout segmentation framework, which while being trained on only one labeled page per manuscript still achieves state-of-the-art performance compared to other popular approaches trained on all the available data when tested on a challenging dataset of ancient Arabic manuscripts.

****

Do We Still Need Non-Maximum Suppression? Accurate Confidence Estimates and Implicit Duplication Modeling With IoU-Aware Calibration

Johannes Gilg, Torben Teepe, Fabian Herzog, Philipp Wolters, Gerhard Rigoll; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4850-4859

Object detectors are at the heart of many semi- and fully autonomous decision systems and are poised to become even more indispensable. They are, however, still lacking in accessibility and can sometimes produce unreliable predictions. Especially concerning in this regard are the (essentially hand-crafted) non-maximum suppression algorithms that lead to an obfuscated prediction process and biased confidence estimates. We show that we can eliminate classic NMS-style post-processing by using IoU-aware calibration. IoU-aware calibration is a conditional Beta calibration; this makes it parallelizable with no hyper-parameters. Instead of arbitrary cutoffs or discounts, it implicitly accounts for the likelihood of each detection being a duplicate and adjusts the confidence score accordingly, resulting in empirically based precision estimates for each detection. Our extensive experiments on diverse detection architectures show that the proposed IoU-aware calibration can successfully model duplicate detections and improve calibration. Compared to the standard sequential NMS and calibration approach, our joint modeling can deliver performance gains over the best NMS-based alternative while producing consistently better-calibrated confidence predictions with less complexity.

****

On the Importance of Large Objects in CNN Based Object Detection Algorithms

Ahmed Ben Saad, Gabriele Facciolo, Axel Davy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 533-542

Object detection models, a prominent class of machine learning algorithms, aim to identify and precisely locate objects in images or videos. However, the task of accurately localizing objects within images yields uneven performances sometimes caused by the objects sizes and the quality of the images and labels. In this paper, we highlight the importance of large objects in learning features that are critical for all sizes. Given these findings, we propose to address this by introducing a weighting term into the loss during training. This term is a function of the object area size. We show that giving more weight to large objects leads to improvement in detection scores across all sizes and so an overall improvement in Object Detectors performances (+2% mAP on small objects, +2% on medium and +4% on large on COCO val 2017 with InternImage-T). Additional experiments and ablation studies with different models and on different dataset further confirm the robustness of our findings.

****

Learning Intra-Class Multimodal Distributions With Orthonormal Matrices

Jumpei Goto, Yohei Nakata, Kiyofumi Abe, Yasunori Ishii, Takayoshi Yamashita; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1870-1879

In this paper, we address the challenges of representing feature distributions which have multimodality within a class in deep neural networks. Existing online clustering methods employ sub-centroids to capture intra-class variations. However, conducting online clustering faces some limitations, i.e., online clustering assigns only a single subcentroid to a feature vector extracted from a backbone and ignores the relationship between the other sub-centroids and the feature vector, and updating sub-centroids in an online clustering manner incurs significant storage costs. To address these limitations, we propose a novel method utilizing orthonormal matrices instead of sub-centroids for relaxing discrete assignme nts into continuous assignments. We update the orthonormal matrices using a gradient-based method, which eliminates the need for online clustering or additional storage. Experimental results on the CIFAR and ImageNet datasets exhibit that the proposed method outperforms current online clustering techniques in classification accuracy, sub-category discovery, and transferability, providing an efficient solution to the challenges posed by complex recognition targets.

****

Assessing Neural Network Robustness via Adversarial Pivotal Tuning

Peter Ebert Christensen, Vésteinn Snæbjarnarson, Andrea Dittadi, Serge Belongie, Sagie Benaim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2952-2961

The robustness of image classifiers is essential to their deployment in the real world. The ability to assess this resilience to manipulations or deviations from the training data is thus crucial. These modifications have traditionally consisted of minimal changes that still manage to fool classifiers, and modern approaches are increasingly robust to them. Semantic manipulations that modify elements of an image in meaningful ways have thus gained traction for this purpose. However, they have primarily been limited to style, color, or attribute changes. While expressive, these manipulations do not make use of the full capabilities of a pretrained generative model. In this work, we aim to bridge this gap. We show how a pretrained image generator can be used to semantically manipulate images in a detailed, diverse, and photorealistic way while still preserving the class of the original image. Inspired by recent GAN-based image inversion methods, we propose a method called Adversarial Pivotal Tuning (APT). Given an image, APT first finds a pivot latent space input that reconstructs the image using a pretrained generator. It then adjusts the generator's weights to create small yet semantic manipulations in order to fool a pretrained classifier. APT preserves the full expressive editing capabilities of the generative model. We demonstrate that APT is capable of a wide range of class-preserving semantic image manipulations that fool a variety of pretrained classifiers. Finally, we show that classifiers that are robust to other benchmarks are not robust to APT manipulations and suggest a method to improve them.

****

Opinion Unaware Image Quality Assessment via Adversarial Convolutional Variational Autoencoder

Ankit Shukla, Avinash Upadhyay, Swati Bhugra, Manoj Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2153-2163

Image quality assessment is a challenging computer vision task due to the lack of corresponding reference (pristine) images. This no-reference bottleneck has been tackled with the utilisation of subjective mean opinion scores (MOS) termed as supervised blind image quality assessment (BIQA) methods. However, inaccessible opinion score scenarios limits their applicability. To relieve these limitations, we propose to employ reconstruction based learning trained only on pristine images. This permits an implicit distribution learning of pristine images and the deviation from this learned feature distribution is subsequently utilised for unsupervised image quality assessment. Specifically, an adversarial convolutional variational auto-encoder framework is employed with KL divergence, perceptual and discriminator loss. With state-of-the-art results on four benchmark datasets, we demonstrate the effectiveness of our proposed framework. An ablation study has also been conducted to highlight the contribution of each module i.e. loss and quality metric for an efficient unsupervised BIQA.

****

A Geometry Loss Combination for 3D Human Pose Estimation

Ai Matsune, Shichen Hu, Guangquan Li, Sihan Wen, Xiantan Zhu, Zhiming Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3272-3281

Root-relative loss has formed the basis of 3D human pose estimation for many years. However, this point-to-point loss treats every keypoint separately and ignores internal connection information of the human body. This leads to illegal pose prediction, which humans cannot form in the real world. It also suffers from differences in estimation difficulty between keypoints. The farther the keypoint is from the torso, the less accurate it is. To address the above problems, this paper proposes geometry loss combination to utilize the geometric relationship between each keypoint fully. This loss combination consists of three loss functions: root-relative pose, bone length, and body part orientation. The previous two have already been used in prior works. Beyond them, we further develop a loss function called body part orientation loss for local body parts. Intuitively, the human body can be divided into three parts: the head, torso, and limbs. Based on this, we select the corresponding keypoints and create virtual planes for each body part. Experiments with different datasets and models demonstrate that our proposed method improves the prediction accuracy. We also achieve MPJPE of 65.0 on the 3DPW test set, which outperforms state-of-the-art methods.

****

Few-Shot Generative Model for Skeleton-Based Human Action Synthesis Using Cross-Domain Adversarial Learning

Kenichiro Fukushi, Yoshitaka Nozaki, Kosuke Nishihara, Kentaro Nakahara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3946-3955

We propose few-shot generative models of skeleton-based human actions on limited samples of the target domain. We exploit large public datasets as a source of motion variations by introducing novel cross-domain and entropy regularization losses that effectively transfer the diversity of the motions contained in the source to the target domain. First, target samples are divided into patches, which are a set of short motion clips. For each patch, we search for a reference motion from the source dataset that is similar to the patch. Next, in adversarial training, our cross-domain regularization encourages the generated sequences to resemble the reference motion at the patch level. Entropy regularization prevents mode collapse by forcing the generator to follow the distribution of the source dataset. Experiments are performed on public datasets where we utilize three action classes from NTU RGB+D 120 as the target and all data of 60 action classes in NTU RGB+D as the source. Ten samples for each target action class, 30 in total, are selected as target data. The results demonstrate that data augmented with the proposed method improve recognition accuracy by 28 % using a ST-GCN classifier.

****

Linking Convolutional Kernel Size to Generalization Bias in Face Analysis CNNs

Hao Liang, Josue Ortega Caro, Vikram Maheshri, Ankit B. Patel, Guha Balakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4705-4715

Training dataset biases are by far the most scrutinized factors when explaining algorithmic biases of neural networks. In contrast, hyperparameters related to the neural network architecture have largely been ignored even though different network parameterizations are known to induce different implicit biases over learned features. For example, convolutional kernel size is known to affect the frequency content of features learned in CNNs. In this work, we present a causal framework for linking an architectural hyperparameter to out-of-distribution algorithmic bias. Our framework is experimental, in that we train several versions of a network with an intervention to a specific hyperparameter, and measure the resulting causal effect of this choice on performance bias when a particular out-of-distribution image perturbation is applied. In our experiments, we focused on measuring the causal relationship between convolutional kernel size and face analysis classification bias across different subpopulations (race/gender), with respect to high-frequency image details. We show that modifying kernel size, even in one layer of a CNN, changes the frequency content of learned features significantly across data subgroups leading to biased generalization performance even in the presence of a balanced dataset.

****

Cross-Attention Between Satellite and Ground Views for Enhanced Fine-Grained Robot Geo-Localization

Dong Yuan, Frederic Maire, Feras Dayoub; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1249-1256

Cross-view image geo-localization aims to determine the locations of outdoor robots by mapping current street-view images with GPS-tagged satellite image patches. Recent works have attained a remarkable level of accuracy in identifying which satellite patches the robot is in, where the location of the central pixel within the matched satellite patch is used as the robot coarse location estimation. This work focuses on robot fine-grained localization within a known satellite patch. Existing fine-grain localization work utilizes correlation operation to obtain similarity between satellite image local descriptors and street-view global descriptors. The correlation operation based on liner matching simplifies the interaction process between two views, leading to a large distance error and affecting model generalization. To address this issue, we devise a cross-view feature fusion network with self-attention and cross-attention layers to replace correlation operation. Additionally, we combine classification and regression prediction to further decrease location distance error. Experiments show that our novel network architecture outperforms the state-of-the-art, exhibiting better generalization capabilities in unseen areas. Specifically, our method reduces the median localization distance error by 43% and 50% respectively in the same area and unseen areas on the VIGOR benchmark.

****

StyleGAN-Fusion: Diffusion Guided Domain Adaptation of Image Generators

Kunpeng Song, Ligong Han, Bingchen Liu, Dimitris Metaxas, Ahmed Elgammal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5453-5463

Can a text-to-image diffusion model be used as a training objective for adapting a GAN generator to another domain? In this paper, we show that the classifier-free guidance can be leveraged as a critic and enable generators to distill knowledge from large-scale text-to-image diffusion models. Generators can be efficiently shifted into new domains indicated by text prompts without access to ground truth samples from target domains. We demonstrate the effectiveness and controllability of our method through extensive experiments. Although not trained to minimize CLIP loss, our model achieves equally high CLIP scores and significantly lower FID than prior work on short prompts and outperforms the baseline qualitatively and quantitatively on long and complicated prompts. To our best knowledge, the proposed method is the first attempt at incorporating large-scale pre-trained diffusion models and distillation sampling for text-driven image generator domain adaptation and gives a quality previously beyond possible. Moreover, we extend our work to 3D-aware style-based generators and DreamBooth guidance.

****

TSP-Transformer: Task-Specific Prompts Boosted Transformer for Holistic Scene Understanding

Shuo Wang, Jing Li, Zibo Zhao, Dongze Lian, Binbin Huang, Xiaomei Wang, Zhengxin Li, Shenghua Gao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 925-934

Holistic scene understanding includes semantic segmentation, surface normal estimation, object boundary detection, depth estimation, etc. The key aspect of this problem is to learn representation effectively, as each subtask builds upon not only correlated but also distinct attributes. Inspired by visual-prompt tuning, we propose a Task-Specific Prompts Transformer, dubbed TSP-Transformer, for holistic scene understanding. It features a vanilla transformer in the early stage and tasks-specific prompts transformer encoder in the lateral stage, where tasks-specific prompts are augmented. By doing so, the transformer layer learns the generic information from the shared parts and is endowed with task-specific capacity. First, the tasks-specific prompts serve as induced priors for each task effectively. Moreover, the task-specific prompts can be seen as switches to favor task-specific representation learning for different tasks. Extensive experiments on NYUD-v2 and PASCAL-Context show that our method achieves state-of-the-art performance, validating the effectiveness of our method for holistic scene understanding.

****

Late to the Party? On-Demand Unlabeled Personalized Federated Learning

Ohad Amosy, Gal Eyal, Gal Chechik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2184-2193

In Federated Learning (FL), multiple clients collaborate to learn a shared model through a central server while keeping data decentralized. Personalized Federated Learning (PFL) further extends FL by learning a personalized model per client. In both FL and PFL, all clients participate in the training process and their labeled data are used for training. However, in reality, novel clients may wish to join a prediction service after it has been deployed, obtaining predictions for their own unlabeled data. Here, we introduce a new learning setup, On-Demand Unlabeled PFL (OD-PFL), where a system trained on a set of clients, needs to be later applied to novel unlabeled clients at inference time. We propose a novel approach to this problem, ODPFL-HN, which learns to produce a new model for the late-to-the-party client. Specifically, we train an encoder network that learns a representation for a client given its unlabeled data. That client representation is fed to a hypernetwork that generates a personalized model for that client. Evaluated on five benchmark datasets, we find that ODPFL-HN generalizes better than the current FL and PFL methods, especially when the novel client has a large shift from training clients. We also analyzed the generalization error for novel clients, and showed analytically and experimentally how novel clients can apply differential privacy to protect their data.

****

EfficientAD: Accurate Visual Anomaly Detection at Millisecond-Level Latencies

Kilian Batzner, Lars Heckler, Rebecca König; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 128-138

Detecting anomalies in images is an important task, especially in real-time computer vision applications. In this work, we focus on computational efficiency and propose a lightweight feature extractor that processes an image in less than a millisecond on a modern GPU. We then use a student-teacher approach to detect anomalous features. We train a student network to predict the extracted features of normal, i.e., anomaly-free training images. The detection of anomalies at test time is enabled by the student failing to predict their features. We propose a training loss that hinders the student from imitating the teacher feature extractor beyond the normal images. It allows us to drastically reduce the computational cost of the student-teacher model, while improving the detection of anomalous features. We furthermore address the detection of challenging logical anomalies that involve invalid combinations of normal local features, for example, a wrong ordering of objects. We detect these anomalies by efficiently incorporating an autoencoder that analyzes images globally. We evaluate our method, called EfficientAD, on 32 datasets from three industrial anomaly detection dataset collections. EfficientAD sets new standards for both the detection and the localization of anomalies. At a latency of two milliseconds and a throughput of six hundred images per second, it enables a fast handling of anomalies. Together with its low error rate, this makes it an economical solution for real-world applications and a fruitful basis for future research.

****

Implicit Neural Representation for Change Detection

Peter Naylor, Diego Di Carlo, Arianna Traviglia, Makoto Yamada, Marco Fiorucci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 935-945

Identifying changes in a pair of 3D aerial LiDAR point clouds, obtained during two distinct time periods over the same geographic region presents a significant challenge due to the disparities in spatial coverage and the presence of noise in the acquisition system. The most commonly used approaches to detecting changes in point clouds are based on supervised methods which necessitate extensive labelled data often unavailable in real-world applications. To address these issues, we propose an unsupervised approach that comprises two components: Implicit Neural Representation (INR) for continuous shape reconstruction and a Gaussian Mixture Model for categorising changes. INR offers a grid-agnostic representation f or encoding bi-temporal point clouds, with unmatched spatial support that can be regularised to enhance high-frequency details and reduce noise. The reconstructions at each timestamp are compared at arbitrary spatial scales, leading to a significant increase in detection capabilities. We apply our method to a benchmark dataset comprising simulated LiDAR point clouds for urban sprawling. This dataset encompasses diverse challenging scenarios, varying in resolutions, input modalities and noise levels. This enables a comprehensive multi-scenario evaluation, comparing our method with the current state-of-the-art approach. We outperform the previous methods by a margin of 10% in the intersection over union metric. In addition, we put our techniques to practical use by applying them in a real-world scenario to identify instances of illicit excavation of archaeological sites and validate our results by comparing them with findings from field experts.

****

Maximum Knowledge Orthogonality Reconstruction With Gradients in Federated Learning

Feng Wang, Senem Velipasalar, M. Cenk Gursoy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3884-3893

Federated learning (FL) aims at keeping client data local to preserve privacy. Instead of gathering the data itself, the server only collects aggregated gradient updates from clients. Following the popularity of FL, there has been considerable amount of work, revealing the vulnerability of FL approaches by reconstructing the input data from gradient updates. Yet, most existing works assume an FL setting with unrealistically small batch size, and have poor image quality when the batch size is large. Other works modify the neural network architectures or parameters to the point of being suspicious, and thus, can be detected by clients. Moreover, most of them can only reconstruct one sample input from a large batch. To address these limitations, we propose a novel and completely analytical approach, referred to as the maximum knowledge orthogonality reconstruction (MKOR), to reconstruct clients' input data. Our proposed method reconstructs a mathematically proven high quality image from large batches. MKOR only requires the server to send secretly modified parameters to clients and can efficiently and inconspicuously reconstruct the input images from clients' gradient updates. We evaluate MKOR's performance on the MNIST, CIFAR-100, and ImageNet dataset and compare it with the state-of-the-art works. The results show that MKOR outperforms the existing approaches, and draws attention to a pressing need for further research on the privacy protection of FL so that comprehensive defense approaches can be developed.

****

ENIGMA-51: Towards a Fine-Grained Understanding of Human Behavior in Industrial Scenarios

Francesco Ragusa, Rosario Leonardi, Michele Mazzamuto, Claudia Bonanno, Rosario Scavo, Antonino Furnari, Giovanni Maria Farinella; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4549-4559

ENIGMA-51 is a new egocentric dataset acquired in an industrial scenario by 19 subjects who followed instructions to complete the repair of electrical boards using industrial tools (e.g., electric screwdriver) and equipments (e.g., oscilloscope). The 51 egocentric video sequences are densely annotated with a rich set of labels that enable the systematic study of human behavior in the industrial domain. We provide benchmarks on four tasks related to human behavior: 1) untrimmed temporal detection of human-object interactions, 2) egocentric human-object interaction detection, 3) short-term object interaction anticipation and 4) natural language understanding of intents and entities. Baseline results show that the ENIGMA-51 dataset poses a challenging benchmark to study human behavior in industrial scenarios. We publicly release the dataset at https://iplab.dmi.unict.it/ENIGMA-51.

****

HELA-VFA: A Hellinger Distance-Attention-Based Feature Aggregation Network for Few-Shot Classification

Gao Yu Lee, Tanmoy Dam, Daniel Puiu Poenar, Vu N. Duong, Md Meftahul Ferdaus; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (

WACV), 2024, pp. 2173-2183

Enabling effective learning using only a few presented examples is a crucial but difficult computer vision objective. Few-shot learning have been proposed to address the challenges, and more recently variational inference-based approaches are incorporated to enhance few-shot classification performances. However, the current dominant strategy utilized the Kullback-Leibler (KL) divergences to find the log marginal likelihood of the target class distribution, while neglecting the possibility of other probabilistic comparative measures, as well as the possibility of incorporating attention in the feature extraction stages, which can increase the effectiveness of the few-shot model. To this end, we proposed the HELlinger-Attention Variational Feature Aggregation network (HELA-VFA), which utilized the Hellinger distance along with attention in the encoder to fulfill the aforementioned gaps. We show that our approach enables the derivation of an alternate form of the lower bound commonly presented in prior works, thus making the variational optimization feasible and be trained on the same footing in a given setting. Extensive experiments performed on four benchmarked few-shot classification datasets demonstrated the feasibility and superiority of our approach relative to the State-Of-The-Arts (SOTAs) approaches.

****

ScanEnts3D: Exploiting Phrase-to-3D-Object Correspondences for Improved Visio-Linguistic Models in 3D Scenes

Ahmed Abdelreheem, Kyle Olszewski, Hsin-Ying Lee, Peter Wonka, Panos Achlioptas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3524-3534

The two popular datasets ScanRefer [20] and ReferIt3D [5] connect natural language to real-world 3D scenes. In this paper, we curate a complementary dataset extending both the aforementioned ones. We associate all objects mentioned in a referential sentence with their underlying instances inside a 3D scene. In contrast, previous work did this only for a single object per sentence. Our Scan Entities in 3D (ScanEnts3D) dataset provides explicit cor- respondences between 369k objects across 84k referential sentences, covering 705 real-world scenes. We propose novel architecture modifications and losses that enable learning from this new type of data and improve the performance for both neural listening and language generation. For neu- ral listening, we improve the SoTA in both the Nr3D and ScanRefer benchmarks by 4.3% and 5.0%, respectively. For language generation, we improve the SoTA by 13.2 CIDEr points on the Nr3D benchmark. For both of these tasks, the new type of data is only used to improve training, but no additional annotations are required at inference time. Our introduced dataset is available on the project's webpage at https://scanents3d.github.io/.

****

A Closer Look at Robustness of Vision Transformers to Backdoor Attacks

Akshayvarun Subramanya, Soroush Abbasi Koohpayegani, Aniruddha Saha, Ajinkya Tejankar, Hamed Pirsiavash; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3874-3883

Transformer architectures are based on self-attention mechanism that processes images as a sequence of patches. As their design is quite different compared to CNNs, it is important to take a closer look at their vulnerability to backdoor attacks and how different transformer architectures affect robustness. Backdoor attacks happen when an attacker poisons a small part of the training images with a specific trigger or backdoor which will be activated later. The model performance is good on clean test images, but the attacker can manipulate the decision of the model by showing the trigger on an image at test time. In this paper, we compare state-of-the-art architectures through the lens of backdoor attacks, specifically how attention mechanisms affect robustness. We observe that the well known vision transformer architecture (ViT) is the least robust architecture and ResMLP, which belongs to a class called Feed Forward Networks (FFN), is most robust to backdoor attacks among state-of-the-art architectures. We also find an intriguing difference between transformers and CNNs -- interpretation algorithms effectively highlight the trigger on test images for transformers but not for CNNs. Based on this observation, we find that a test-time image blocking defense redu ces the attack success rate by a large margin for transformers. We also show that such blocking mechanisms can be incorporated during the training process to improve robustness even further. We believe our experimental findings will encourage the community to understand the building block components in developing novel architectures robust to backdoor attacks. Code is available here:https://github.com/UCDvision/backdoor_transformer.git

****

Differentiable JPEG: The Devil Is in the Details

Christoph Reich, Biplob Debnath, Deep Patel, Srimat Chakradhar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4126-4135

JPEG remains one of the most widespread lossy image coding methods. However, the non-differentiable nature of JPEG restricts the application in deep learning pipelines. Several differentiable approximations of JPEG have recently been proposed to address this issue. This paper conducts a comprehensive review of existing diff. JPEG approaches and identifies critical details that have been missed by previous methods. To this end, we propose a novel diff. JPEG approach, overcoming previous limitations. Our approach is differentiable w.r.t. the input image, the JPEG quality, the quantization tables, and the color conversion parameters. We evaluate the forward and backward performance of our diff. JPEG approach against existing methods. Additionally, extensive ablations are performed to evaluate crucial design choices. Our proposed diff. JPEG resembles the (non-diff.) reference implementation best, significantly surpassing the recent-best diff. approach by 3.47dB (PSNR) on average. For strong compression rates, we can even improve PSNR by 9.51dB. Strong adversarial attack results are yielded by our diff. JPEG, demonstrating the effective gradient approximation. Our code is available at https://github.com/necla-ml/Diff-JPEG.

****

CLIP-DIY: CLIP Dense Inference Yields Open-Vocabulary Semantic Segmentation For-Free

Monika Wysocza■ska, Michaël Ramamonjisoa, Tomasz Trzci■ski, Oriane Siméoni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1403-1413

The emergence of CLIP has opened the way for open-world image perception. The zero-shot classification capabilities of the model are impressive but are harder to use for dense tasks such as image segmentation. Several methods have proposed different modifications and learning schemes to produce dense output. Instead, we propose in this work an open-vocabulary semantic segmentation method, dubbed CLIP-DIY, which does not require any additional training or annotations, but instead leverages existing unsupervised object localization approaches. In particular, CLIP-DIY is a multi-scale approach that directly exploits CLIP classification abilities on patches of different sizes and aggregates the decision in a single map. We further guide the segmentation using foreground/background scores obtained using unsupervised object localization methods. With our method, we obtain state-of-the-art zero-shot semantic segmentation results on PASCAL VOC and perform on par with the best methods on COCO.

****

Dual Domain Diffusion Guidance for 3D CBCT Metal Artifact Reduction

Yongjin Choi, Doeyoung Kwon, Seung Jun Baek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7965-7974

Previous methods to solve the problem of metal artifact reduction (MAR) have mostly focused on 2D MAR, making it challenging to apply to problems with 3-dimensional CT such as CBCT. In this paper, we propose a novel approach for 3D MAR which utilizes two diffusion models to model the metal-free CBCT prior and metal artifact prior. Through dual-domain guidance in the image and projection domains, the 3D connectivity is enhanced in the restored images. Moreover, we propose a memory-efficient technique for an efficient sampling of 3-dimensional data, which reduces the memory usage by orders of magnitude. Experiments show that our method achieves the state-of-the-art performance not only with synthetic data but also with real-world clinical and out-of-distribution data.

****

Joint 3D Shape and Motion Estimation From Rolling Shutter Light-Field Images

Hermès McGriff, Renato Martins, Nicolas Andreff, Cédric Demonceaux; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3699-3708

In this paper, we propose an approach to address the problem of 3D reconstruction of scenes from a single image captured by a light-field camera equipped with a rolling shutter sensor. Our method leverages the 3D information cues present in the light-field and the motion information provided by the rolling shutter effect. We present a generic model for the imaging process of this sensor and a two-stage algorithm that minimizes the re-projection error while considering the position and motion of the camera in a motion-shape bundle adjustment estimation strategy. Thereby, we provide an instantaneous 3D shape-and-pose-and-velocity sensing paradigm. To the best of our knowledge, this is the first study to leverage this type of sensor for this purpose. We also present a new benchmark dataset composed of different light-fields showing rolling shutter effects, which can be used as a common base to improve the evaluation and tracking the progress in the field. We demonstrate the effectiveness and advantages of our approach through several experiments conducted for different scenes and types of motions. The source code and dataset are publicly available at: https://github.com/ICB-Vision-AI/RSLF

****

ConeQuest: A Benchmark for Cone Segmentation on Mars

Mirali Purohit, Jacob Adler, Hannah Kerner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6026-6035

Over the years, space scientists have collected terabytes of Mars data from satellites and rovers. One important set of features identified in Mars orbital images is pitted cones, which are interpreted to be mud volcanoes believed to form in regions that were once saturated in water (i.e., a lake or ocean). Identifying pitted cones globally on Mars would be of great importance, but expert geologists are unable to sort through the massive orbital image archives to identify all examples. However, this task is well suited for computer vision. Although several computer vision datasets exist for various Mars-related tasks, there is currently no open-source dataset available for cone detection/segmentation. Furthermore, previous studies trained models using data from a single region, which limits their applicability for global detection and mapping. Motivated by this, we introduce ConeQuest, the first expert-annotated public dataset to identify cones on Mars. ConeQuest consists of >13k samples from 3 different regions of Mars. We propose two benchmark tasks using ConeQuest: (i) Spatial Generalization and (ii) Cone-size Generalization. We finetune and evaluate widely-used segmentation models on both benchmark tasks. Results indicate that cone segmentation is a challenging open problem not solved by existing segmentation models, which achieve an average IoU of 52.52% and 42.55% on in-distribution data for tasks (i) and (ii), respectively. We believe this new benchmark dataset will facilitate the development of more accurate and robust models for cone segmentation. Data and code are available at https://github.com/kerner-lab/ConeQuest.

****

A Multimodal Benchmark and Improved Architecture for Zero Shot Learning

Keval Doshi, Amanmeet Garg, Burak Uzkent, Xiaolong Wang, Mohamed Omar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2021-2030

In this work, we demonstrate that due to the inadequacies in the existing evaluation protocols and datasets, there is a need to revisit and comprehensively examine the multimodal Zero-Shot Learning (MZSL) problem formulation. Specifically, we address two major challenges faced by current MZSL approaches; (1) Established baselines are frequently incomparable and occasionally even flawed since existing evaluation datasets often have some overlap with the training dataset, thus violating the zero-shot paradigm; (2) Most existing methods are biased towards seen classes, which significantly reduces the performance when evaluated on both seen and unseen classes. To address these challenges, we first introduce a new m ultimodal dataset for zero-shot evaluation called MZSL-50 with 4462 videos from 50 widely diversified classes and no overlap with the training data. Further, we propose a novel multimodal zeroshot transformer (MZST) architecture that leverages attention bottlenecks for multimodal fusion. Our model directly predicts the semantic representation and is superior at reducing the bias towards seen classes. We conduct extensive ablation studies, and achieve state-of-the-art results on three benchmark datasets and our novel MZSL-50 dataset. Specifically, we improve the conventional MZSL performance by a margin of 2.1%, 9.81% and 8.68% on VGGSound, UCF-101 and ActivityNet, respectively. Finally, we expect the introduction of the MZSL-50 dataset will promote the future in-depth research on multimodal zero-shot learning in the community.

****

PlantPlotGAN: A Physics-Informed Generative Adversarial Network for Plant Disease Prediction

Felipe A. Lopes, Vasit Sagan, Flavio Esposito; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7066-7075

Monitoring plantations is crucial for crop management and producing healthy harvests. Unmanned Aerial Vehicles (UAVs) have been used to collect multispectral images that aid in this monitoring. However, given the number of hectares to be monitored and the limitations of flight, plant disease signals become visually clear only in the later stages of plant growth and only if the disease has spread throughout a significant portion of the plantation. This limited amount of relevant data hampers the prediction models, as the algorithms struggle to generalize patterns with unbalanced or unrealistic augmented datasets effectively. To address this issue, we propose PlantPlotGAN, a physics-informed generative model capable of reproducing synthetic multispectral plot images with realistic vegetation indices. These indices served as a proxy for early disease detection and were used to evaluate if our model could help increase the accuracy of prediction models. The results demonstrate that the synthetic imagery generated from PlantPlotGAN outperforms state-of-the-art methods regarding the Frechet inception distance. Moreover, prediction models achieve higher accuracy metrics when trained with synthetic and original imagery for earlier plant disease detection compared to the training processes based solely on real imagery.

****

Common Diffusion Noise Schedules and Sample Steps Are Flawed

Shanchuan Lin, Bingchen Liu, Jiashi Li, Xiao Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5404-5411

We discover that common diffusion noise schedules do not enforce the last timestep to have zero signal-to-noise ratio (SNR), and some implementations of diffusion samplers do not start from the last timestep. Such designs are flawed and do not reflect the fact that the model is given pure Gaussian noise at inference, creating a discrepancy between training and inference. We show that the flawed design causes real problems in existing implementations. In Stable Diffusion, it severely limits the model to only generate images with medium brightness and prevents it from generating very bright and dark samples. We propose a few simple fixes: (1) rescale the noise schedule to enforce zero terminal SNR; (2) train the model with v prediction; (3) change the sampler to always start from the last timestep; (4) rescale classifier-free guidance to prevent over-exposure. These simple changes ensure the diffusion process is congruent between training and inference and allow the model to generate samples more faithful to the original data distribution.

****

Efficient Expansion and Gradient Based Task Inference for Replay Free Incremental Learning

Soumya Roy, Vinay Verma, Deepak Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1165-1175

This paper proposes a simple but highly efficient expansion-based model for continual learning. The recent feature transformation, masking and factorization-based methods are efficient, but they grow the model only over the global or shared parameter. Therefore, these approaches do not fully utilize the previously lear ned information because the same task-specific parameter forgets the earlier knowledge. Thus, these approaches show limited transfer learning ability. Moreover, most of these models have constant parameter growth for all tasks, irrespective of the task complexity. Our work proposes a simple filter and channel expansion-based method that grows the model over the previous task parameters and not just over the global parameter. Therefore, it fully utilizes all the previously learned information without forgetting, which results in better knowledge transfer. The growth rate in our proposed model is a function of task complexity; therefore for a simple task, the model has a smaller parameter growth, while for complex tasks, the model requires more parameters to adapt to the current task. Recent expansion-based models show promising results for task incremental learning (TIL). However, for class incremental learning (CIL), prediction of task id is a crucial challenge; hence, their results degrade rapidly as the number of tasks increase. In this work, we propose a robust task prediction method that leverages entropy weighted data augmentations and the model's gradient using pseudo labels. We evaluate our model on various datasets and architectures in the TIL, CIL and generative continual learning settings. The proposed approach shows state-of-the-art results in all these settings. Our extensive ablation studies show the efficacy of the proposed components.

****

PolyMaX: General Dense Prediction With Mask Transformer

Xuan Yang, Liangzhe Yuan, Kimberly Wilber, Astuti Sharma, Xiuye Gu, Siyuan Qiao, Stephanie Debats, Huisheng Wang, Hartwig Adam, Mikhail Sirotenko, Liang-Chieh Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1050-1061

Dense prediction tasks, such as semantic segmentation, depth estimation, and surface normal prediction, can be easily formulated as per-pixel classification (discrete outputs) or regression (continuous outputs). This per-pixel prediction paradigm has remained popular due to the prevalence of fully convolutional networks. However, on the recent frontier of segmentation task, the community has been witnessing a shift of paradigm from per-pixel prediction to cluster-prediction with the emergence of transformer architectures, particularly the mask transformers, which directly predicts a label for a mask instead of a pixel. Despite this shift, methods based on the per-pixel prediction paradigm still dominate the benchmarks on the other dense prediction tasks that require continuous outputs, such as depth estimation and surface normal prediction. Motivated by the success of DORN and AdaBins in depth estimation, achieved by discretizing the continuous output space, we propose to generalize the cluster-prediction based method to general dense prediction tasks. This allows us to unify dense prediction tasks with the mask transformer framework. Remarkably, the resulting model PolyMaX demonstrates state-of-the-art performance on three benchmarks of NYUD-v2 dataset. We hope our simple yet effective design can inspire more research on exploiting mask transformers for more dense prediction tasks. Code and model will be made available.

****

Approximating Intersections and Differences Between Linear Statistical Shape Models Using Markov Chain Monte Carlo

Maximilian Weiherer, Finn Klein, Bernhard Egger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6364-6373

To date, the comparison of Statistical Shape Models (SSMs) is often solely performance-based, carried out by means of simplistic metrics such as compactness, generalization, or specificity. Any similarities or differences between the actual shape spaces can neither be visualized nor quantified. In this paper, we present a new method to qualitatively compare two linear SSMs in dense correspondence by computing approximate intersection spaces and set-theoretic differences between the (hyper-ellipsoidal) allowable shape domains spanned by the models. To this end, we approximate the distribution of shapes lying in the intersection space using Markov chain Monte Carlo and subsequently apply Principal Component Analysis (PCA) to the posterior samples, eventually yielding a new SSM of the intersection space. We estimate differences between linear SSMs in a similar manner; he re, however, the resulting spaces are no longer convex and we do not apply PCA but instead use the posterior samples for visualization. We showcase the proposed algorithm qualitatively by computing and analyzing intersection spaces and differences between publicly available face models, focusing on gender-specific male and female as well as identity and expression models. Our quantitative evaluation based on SSMs built from synthetic and real-world data sets provides detailed evidence that the introduced method is able to recover ground-truth intersection spaces and differences accurately.

****

Few-Shot Shape Recognition by Learning Deep Shape-Aware Features

Wenlong Shi, Changsheng Lu, Ming Shao, Yinjie Zhang, Siyu Xia, Piotr Koniusz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1848-1859

Traditional shape descriptors have been gradually replaced by convolutional neural networks due to their superior performance in feature extraction and classification. The state-of-the-art methods recognize object shapes via image reconstruction or pixel classification. However, these methods are biased toward texture information and overlook the essential shape descriptions, thus, they fail to generalize to unseen shapes. We are the first to propose a few-shot shape descriptor (FSSD) to recognize object shapes given only one or a few samples. We employ an embedding module for FSSD to extract transformation-invariant shape features. Secondly, we develop a dual attention mechanism to decompose and reconstruct the shape features via learnable shape primitives. In this way, any shape can be formed through a finite set basis, and the learned representation model is highly interpretable and extendable to unseen shapes. Thirdly, we propose a decoding module to include the supervision of shape masks and edges and align the original and reconstructed shape features, enforcing the learned features to be more shape-aware. Lastly, all the proposed modules are assembled into a few-shot shape recognition scheme. Experiments on five datasets show that our FSSD significantly improves the shape classification compared to the state-of-the-art under the few-shot setting.

****

Multi-Class Segmentation From Aerial Views Using Recursive Noise Diffusion

Benedikt Kolbeinsson, Krystian Mikolajczyk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8439-8449

Semantic segmentation from aerial views is a crucial task for autonomous drones, as they rely on precise and accurate segmentation to navigate safely and efficiently. However, aerial images present unique challenges such as diverse viewpoints, extreme scale variations, and high scene complexity. In this paper, we propose an end-to-end multi-class semantic segmentation diffusion model that addresses these challenges. We introduce recursive denoising to allow information to propagate through the denoising process, as well as a hierarchical multi-scale approach that complements the diffusion process. Our method achieves promising results on the UAVid dataset and state-of-the-art performance on the Vaihingen Building segmentation benchmark. Being the first iteration of this method, it shows great promise for future improvements. Our code and models are available at: https://github.com/benediktkol/recursive-noise-diffusion

****

Enhancing Multi-View Pedestrian Detection Through Generalized 3D Feature Pulling

Sithu Aung, Haesol Park, Hyungjoo Jung, Junghyun Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1196-1205

The main challenge in multi-view pedestrian detection is integrating view-specific features into a unified space for comprehensive end-to-end perception. Prior multi-view detection methods have focused on projecting perspective-view features onto the ground plane, creating a "bird's eye view" (BEV) representation of the scene. This paper proposes a simple but effective architecture that utilizes a non-parametric 3D feature-pulling strategy. This strategy directly extracts the corresponding 2D features for each valid voxel within the 3D feature volume, addressing the feature loss that may arise in previous methods. The proposed frame work introduces three novel modules, each crafted to bolster the generalization capabilities of multi-view detection systems. Through extensive experiments, the efficacy of the proposed model is demonstrated. The results show a new state-of-the-art accuracy, both in conventional scenarios and particularly in the context of scene generalization benchmarks.

****

Automated Sperm Assessment Framework and Neural Network Specialized for Sperm Video Recognition

Takuro Fujii, Hayato Nakagawa, Teppei Takeshima, Yasushi Yumura, Tomoki Hamagami; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7675-7684

Infertility is a global health problem, and an increasing number of couples are seeking medical assistance to achieve reproduction, at least half of which are caused by men. The success rate of assisted reproductive technologies depends on sperm assessment, in which experts determine whether sperm can be used for reproduction based on morphology and motility of sperm. Previous sperm assessment studies with deep learning have used datasets comprising images that include only sperm heads, which cannot consider motility and other morphologies of sperm. Furthermore, the labels of the dataset are one-hot, which provides insufficient support for experts, because assessment results are inconsistent between experts, and they have no absolute answer. Therefore, we constructed the video dataset for sperm assessment whose videos include sperm head as well as neck and tail, and its labels were annotated with soft-label. Furthermore, we proposed the sperm assessment framework and the neural network, RoSTFine, for sperm video recognition. Experimental results showed that RoSTFine could improve the sperm assessment performances compared to existing video recognition models and focus strongly on important sperm parts (i.e., head and neck).

****

Have We Ever Encountered This Before? Retrieving Out-of-Distribution Road Obstacles From Driving Scenes

Youssef Shoeb, Robin Chan, Gesina Schwalbe, Azarm Nowzad, Fatma Güney, Hanno Gottschalk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7396-7406

In the life cycle of highly automated systems operating in an open and dynamic environment, the ability to adjust to emerging challenges is crucial. For systems integrating data-driven AI-based components, rapid responses to deployment issues require fast access to related data for testing and reconfiguration. In the context of automated driving, this especially applies to road obstacles that were not included in the training data, commonly referred to as out-of-distribution (OoD) road obstacles. Given the availability of large uncurated recordings of driving scenes, a pragmatic approach is to query a database to retrieve similar scenarios featuring the same safety concerns due to OoD road obstacles. In this work, we extend beyond identifying OoD road obstacles in video streams and offer a comprehensive approach to extract sequences of OoD road obstacles using text queries, thereby proposing a way of curating a collection of OoD data for subsequent analysis. Our proposed method leverages the recent advances in OoD segmentation and multi-modal foundation models to identify and efficiently extract safety-relevant scenes from unlabeled videos. We present a first approach for the novel task of text-based OoD object retrieval, which addresses the question "Have we ever encountered this before?".

****

Polarimetric PatchMatch Multi-View Stereo

Jinyu Zhao, Jumpei Oishi, Yusuke Monno, Masatoshi Okutomi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3476-3484

PatchMatch Multi-View Stereo (PatchMatch MVS) is one of the popular MVS approaches, owing to its balanced accuracy and efficiency. In this paper, we propose Polarimetric PatchMatch multi-view Stereo (PolarPMS), which is the first method exploiting polarization cues to PatchMatch MVS. The key of PatchMatch MVS is to generate depth and normal hypotheses, which form local 3D planes and slanted stereo matching windows, and efficiently search for the best hypothesis based on the consistency among multi-view images. In addition to standard photometric consistency, our PolarPMS evaluates polarimetric consistency to assess the validness of a depth and normal hypothesis, motivated by the physical property that the polarimetric information is related to the object's surface normal. Experimental results demonstrate that our PolarPMS can improve the accuracy and the completeness of reconstructed 3D models, especially for texture-less surfaces, compared with state-of-the-art PatchMatch MVS methods.

****

High-Fidelity Pseudo-Labels for Boosting Weakly-Supervised Segmentation

Arvi Jonnarth, Yushan Zhang, Michael Felsberg; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1010-1019

Image-level weakly-supervised semantic segmentation (WSSS) reduces the usually vast data annotation cost by surrogate segmentation masks during training. The typical approach involves training an image classification network using global average pooling (GAP) on convolutional feature maps. This enables the estimation of object locations based on class activation maps (CAMs), which identify the importance of image regions. The CAMs are then used to generate pseudo-labels, in the form of segmentation masks, to supervise a segmentation model in the absence of pixel-level ground truth. Our work is based on two techniques for improving CAMs; importance sampling, which is a substitute for GAP, and the feature similarity loss, which utilizes a heuristic that object contours almost always align with color edges in images. However, both are based on the multinomial posterior with softmax, and implicitly assume that classes are mutually exclusive, which turns out suboptimal in our experiments. Thus, we reformulate both techniques based on binomial posteriors of multiple independent binary problems. This has two benefits; their performance is improved and they become more general, resulting in an add-on method that can boost virtually any WSSS method. This is demonstrated on a wide variety of baselines on the PASCAL VOC dataset, improving the region similarity and contour quality of all implemented state-of-the-art methods. Experiments on the MS COCO dataset further show that our proposed add-on is well-suited for large-scale settings. Our code implementation is available at https://github.com/arvijj/hfpl.

****

Optical Flow Domain Adaptation via Target Style Transfer

Jeongbeen Yoon, Sanghyun Kim, Suha Kwak, Minsu Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2111-2121

Optical flows play an integral role for a variety of motion-related tasks such as action recognition, object segmentation, and tracking in videos. While state-of-the-art optical flow methods heavily rely on learning, the learned optical flow methods significantly degrade when applied to different domains, and the training datasets are very limited due to the extreme cost of flow-level annotation. To tackle the issue, we introduce a domain adaptation technique for optical flow estimation. Our method extracts diverse style statistics of the target domain and use them in training to generate synthetic features from the source features, which contain the contents of the source but the style of the target. We also impose motion consistency between the synthetic target and the source and deploy adversarial learning at the flow prediction to encourage domain-invariant features. Experimental results show that the proposed method achieves substantial and consistent improvements in different domain adaptation scenarios on VKITTI 2, Sintel, and KITTI 2015 benchmarks.

****

Controlling Character Motions Without Observable Driving Source

Weiyuan Li, Bin Dai, Ziyi Zhou, Qi Yao, Baoyuan Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6194-6203

How to generate diverse, life-like, and unlimited long head/body sequences without any driving source? We argue that this under-investigated research problem is non-trivial at all, and has unique technical challenges behind it. Without semantic constraints from the driving sources, using the standard autoregressive mod el to generate infinitely long sequences would easily result in 1) out-of-distribution (OOD) issue due to the accumulated error, 2) insufficient diversity to produce natural and life-like motion sequences and 3) undesired periodic patterns along the time. To tackle the above challenges, we propose a systematic framework that marries the benefits of VQ-VAE and a novel token-level control policy trained with reinforcement learning using carefully designed reward functions. A high-level prior model can be easily injected on top to generate unlimited long and diverse sequences. Although we focus on no driving sources now, our framework can be generalized for controlled synthesis with explicit driving sources. Through comprehensive evaluations, we conclude that our proposed framework can address all the above-mentioned challenges and outperform other strong baselines very significantly.

****

Evaluation of Video Masked Autoencoders' Performance and Uncertainty Estimations for Driver Action and Intention Recognition

Koen Vellenga, H. Joe Steinhauer, Göran Falkman, Tomas Björklund; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7429-7437

Traffic fatalities remain among the leading death causes worldwide. To reduce this figure, car safety is listed as one of the most important factors. To actively support human drivers, it is essential for advanced driving assistance systems to be able to recognize the driver's actions and intentions. Prior studies have demonstrated various approaches to recognize driving actions and intentions based on in-cabin and external video footage. Given the performance of self-supervised video pre-trained (SSVP) Video Masked Autoencoders (VMAEs) on multiple action recognition datasets, we evaluate the performance of SSVP VMAEs on the Honda Research Institute Driving Dataset for driver action recognition (DAR) and on the Brain4Cars dataset for driver intention recognition (DIR). Besides the performance, the application of an artificial intelligence system in a safety-critical environment must be capable to express when it is uncertain about the produced results. Therefore, we also analyze uncertainty estimations produced by a Bayes-by-Backprop last-layer (BBB-LL) and Monte-Carlo (MC) dropout variants of an VMAE. Our experiments show that an VMAE achieves a higher overall performance for both offline DAR and end-to-end DIR compared to the state-of-the-art. The analysis of the BBB-LL and MC dropout models show higher uncertainty estimates for incorrectly classified test instances compared to correctly predicted test instances.

****

Nested Diffusion Processes for Anytime Image Generation

Noam Elata, Bahjat Kawar, Tomer Michaeli, Michael Elad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5018-5027

Diffusion models are the current state-of-the-art in image generation, synthesizing high-quality images by breaking down the generation process into many fine-grained denoising steps. Despite their good performance, diffusion models are computationally expensive, requiring many neural function evaluations (NFEs). In this work, we propose an anytime diffusion-based method that can generate viable images when stopped at arbitrary times before completion. Using existing pretrained diffusion models, we show that the generation scheme can be recomposed as two nested diffusion processes, enabling fast iterative refinement of a generated image. In experiments on ImageNet and Stable Diffusion-based text-to-image generation, we show, both qualitatively and quantitatively, that our method's intermediate generation quality greatly exceeds that of the original diffusion model, while the final generation result remains comparable. We illustrate the applicability of Nested Diffusion in several settings, including for solving inverse problems, and for rapid text-based content creation by allowing user intervention throughout the sampling process.

****

Can You Even Tell Left From Right? Presenting a New Challenge for VQA

Sai Raam Venkataraman, Rishi Sridhar Rao, S. Balasubramanian, R. Raghunatha Sarma, Chandra Sekhar Vorugunti; Proceedings of the IEEE/CVF Winter Conference on Ap plications of Computer Vision (WACV), 2024, pp. 4498-4507

Visual Question Answering (VQA) needs a means of evaluating the strengths and weaknesses of models. One aspect of such an evaluation is the measurement of compositional generalisation. This relates to the ability of a model to answer well on scenes whose compositions are different from those of scenes in the training dataset. In this work, we present several quantitative measures of compositional separation and find that popular datasets for VQA are not good compositional evaluators. To solve this, we present Uncommon Objects in Unseen Configurations (UOUC), a synthetic dataset for VQA. UOUC is at once fairly complex while also being compositionally well-separated. The object-class of UOUC consists of 380 clasess taken from 528 characters from the Dungeons and Dragons game. The training dataset of UOUC consists of 200,000 scenes; whereas the test set consists of 30,000 scenes. In order to study compositional generalisation, simple reasoning and memorisation, each scene of UOUC is annotated with up to 10 novel questions. These deal with spatial relationships, hypothetical changes to scenes, counting, comparison, memorisation and memory-based reasoning. In total, UOUC presents over 2 million questions. Our evaluation of recent high-performing models for VQA shows that they exhibit poor compositional generalisation, and comparatively lower ability towards simple reasoning. These results suggest that UOUC could lead to advances in research by being a strong benchmark for VQA, especially in the study of compositional generalisation.

****

2D Feature Distillation for Weakly- and Semi-Supervised 3D Semantic Segmentation

Ozan Unal, Dengxin Dai, Lukas Hoyer, Yigit Baran Can, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7336-7345

As 3D perception problems grow in popularity and the need for large-scale labeled datasets for LiDAR semantic segmentation increase, new methods arise that aim to reduce the necessity for dense annotations by employing weakly-supervised training. However these methods continue to show weak boundary estimation and high false negative rates for small objects and distant sparse regions. We argue that such weaknesses can be compensated by using RGB images which provide a denser representation of the scene. We propose an image-guidance network (IGNet) which builds upon the idea of distilling high level feature information from a domain adapted synthetically trained 2D semantic segmentation network. We further utilize a one-way contrastive learning scheme alongside a novel mixing strategy called FOVMix, to combat the horizontal field-of-view mismatch between the two sensors and enhance the effects of image guidance. IGNet achieves state-of-the-art results for weakly-supervised LiDAR semantic segmentation on ScribbleKITTI, boasting up to 98% relative performance to fully supervised training with only 8% labeled points, while introducing no additional annotation burden or computational/memory cost during inference. Furthermore, we show that our contributions also prove effective for semi-supervised training, where IGNet claims state-of-the-art results on both ScribbleKITTI and SemanticKITTI.

****

MAELi: Masked Autoencoder for Large-Scale LiDAR Point Clouds

Georg Krispel, David Schinagl, Christian Fruhwirth-Reisinger, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3383-3392

The sensing process of large-scale LiDAR point clouds inevitably causes large blind spots, i.e. regions not visible to the sensor. We demonstrate how these inherent sampling properties can be effectively utilized for self-supervised representation learning by designing a highly effective pre-training framework that considerably reduces the need for tedious 3D annotations to train state-of-the-art object detectors. Our Masked AutoEncoder for LiDAR point clouds (MAELi) intuitively leverages the sparsity of LiDAR point clouds in both the encoder and decoder during reconstruction. This results in more expressive and useful initialization, which can be directly applied to downstream perception tasks, such as 3D object detection or semantic segmentation for autonomous driving. In a novel reconstruction approach, MAELi distinguishes between empty and occluded space and emplo ys a new masking strategy that targets the LiDAR's inherent spherical projection. Thereby, without any ground truth whatsoever and trained on single frames only, MAELi obtains an understanding of the underlying 3D scene geometry and semantics. To demonstrate the potential of MAELi, we pre-train backbones in an end-to-end manner and show the effectiveness of our unsupervised pre-trained weights on the tasks of 3D object detection and semantic segmentation.

****

Empowering Unsupervised Domain Adaptation With Large-Scale Pre-Trained Vision-Language Models

Zhengfeng Lai, Haoping Bai, Haotian Zhang, Xianzhi Du, Jiulong Shan, Yinfei Yang, Chen-Nee Chuah, Meng Cao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2691-2701

Unsupervised Domain Adaptation (UDA) aims to leverage the labeled source domain to solve the tasks on the unlabeled target domain. Traditional UDA methods face the challenge of the tradeoff between domain alignment and semantic class discriminability, especially when a large domain gap exists between the source and target domain. The efforts of applying large-scale pre-training to bridge the domain gaps remain limited. In this work, we propose that Vision-Language Models (VLMs) can empower UDA tasks due to their training pattern with language alignment and their large-scale pre-trained datasets. For example, CLIP and GLIP have shown promising zero-shot generalization in classification and detection tasks. However, directly fine-tuning these VLMs into downstream tasks may be computationally expensive and not scalable if we have multiple domains that need to be adapted. Therefore, in this work, we first study an efficient adaption of VLMs to preserve the original knowledge while maximizing its flexibility for learning new knowledge. Then, we design a domain-aware pseudo-labeling scheme tailored to VLMs for domain disentanglement. We show the superiority of the proposed methods in four UDA-classification and two UDA-detection benchmarks, with a significant improvement (+9.9%) on DomainNet.

****

FreMIM: Fourier Transform Meets Masked Image Modeling for Medical Image Segmentation

Wenxuan Wang, Jing Wang, Chen Chen, Jianbo Jiao, Yuanxiu Cai, Shanshan Song, Jiangyun Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7860-7870

The research community has witnessed the powerful potential of self-supervised Masked Image Modeling (MIM), which enables the models capable of learning visual representation from unlabeled data. In this paper, to incorporate both the crucial global structural information and local details for dense prediction tasks, we alter the perspective to the frequency domain and present a new MIM-based framework named FreMIM for self-supervised pre-training to better accomplish medical image segmentation tasks. Based on the observations that the detailed structural information mainly lies in the high-frequency components and the high-level semantics are abundant in the low-frequency counterparts, we further incorporate multi-stage supervision to guide the representation learning during the pre-training phase. Extensive experiments on three benchmark datasets show the superior advantage of our FreMIM over previous state-of-the-art MIM methods. Compared with various baselines trained from scratch, our FreMIM could consistently bring considerable improvements to model performance. The code will be made publicly available at https://github.com/jingw193/FreMIM.

****

Rotation-Constrained Cross-View Feature Fusion for Multi-View Appearance-Based Gaze Estimation

Yoichiro Hisadome, Tianyi Wu, Jiawei Qin, Yusuke Sugano; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5985-5994

Appearance-based gaze estimation has been actively studied in recent years. However, its generalization performance for unseen head poses is still a significant limitation for existing methods. This work proposes a generalizable multi-view gaze estimation task and a cross-view feature fusion method to address this issu

e. In addition to paired images, our method takes the relative rotation matrix between two cameras as additional input. The proposed network learns to extract rotatable feature representation by using relative rotation as a constraint and adaptively fuses the rotatable features via stacked fusion modules. This simple yet efficient approach significantly improves generalization performance under unseen head poses without significantly increasing computational cost. The model can be trained with random combinations of cameras without fixing the positioning and can generalize to unseen camera pairs during inference. Through experiments using multiple datasets, we demonstrate the advantage of the proposed method over baseline methods, including state-of-the-art domain generalization approaches.

****

Continual Atlas-Based Segmentation of Prostate MRI

Amin Ranem, Camila González, Daniel Pinto dos Santos, Andreas M. Bucher, Ahmed E. Othman, Anirban Mukhopadhyay; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7563-7572

Continual learning (CL) methods designed for natural image classification often fail to reach basic quality standards for medical image segmentation. Atlas-based segmentation, a well-established approach in medical imaging, incorporates domain knowledge on the region of interest, leading to semantically coherent predictions. This is especially promising for CL, as it allows us to leverage structural information and strike an optimal balance between model rigidity and plasticity over time. When combined with privacy-preserving prototypes, this process offers the advantages of rehearsal-based CL without compromising patient privacy. We propose Atlas Replay, an atlas-based segmentation approach that uses prototypes to generate high-quality segmentation masks through image registration that maintain consistency even as the training distribution changes. We explore how our proposed method performs compared to state-of-the-art CL methods in terms of knowledge transferability across seven publicly available prostate segmentation datasets. Prostate segmentation plays a vital role in diagnosing prostate cancer, however, it poses challenges due to substantial anatomical variations, benign structural differences in older age groups, and fluctuating acquisition parameters. Our results show that Atlas Replay is both robust and generalizes well to yet-unseen domains while being able to maintain knowledge, unlike end-to-end segmentation methods. Our code base is available under https://github.com/MECLabTUDA/Atlas-Replay.

****

CGAPoseNet+GCAN: A Geometric Clifford Algebra Network for Geometry-Aware Camera Pose Regression

Alberto Pepe, Joan Lasenby, Sven Buchholz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6593-6603

We introduce CGAPoseNet+GCAN, which enhances CGAPoseNet, an architecture for camera pose regression, with a Geometric Clifford Algebra Network (GCAN). With the addition of the GCAN we obtain a geometry-aware pipeline for camera pose regression from RGB images only. CGAPoseNet employs Clifford Geometric Algebra to unify quaternions and translation vectors into a single mathematical object, the motor, which can be used to uniquely describe camera poses. CGAPoseNet solves the issue of balancing rotation and translation components in the loss function, and can obtain comparable results to other approaches without the need of expensive tuning of the loss function or additional information about the scene, such as 3D point clouds, which might not always be available. CGAPoseNet, however, like several approaches in the literature, only learns to predict motor coefficients, and it is unaware of the mathematical space in which predictions sit in and of their geometrical meaning. By leveraging recent advances in Geometric Deep Learning, we modify CGAPoseNet with a GCAN: proposals of possible motor coefficients associated with a camera frame are obtained from the InceptionV3 backbone, and the GCAN downsamples them to a single motor through a sequence of layers that work in G_ 4,0 . The network is hence geometry-aware, has multivector-valued inputs, weights and biases and preserves the grade of the objects that it receives in input. CGAPoseNet+GCAN has almost 4 million fewer trainable parameters, it reduces the average rotation error by 41% and the average translation error by 8.8% compared to CGAPoseNet. Similarly, it reduces rotation and translation errors by 32.6% and 19.9%, respectively, compared to the best performing PoseNet strategy. CGAPoseNet+GCAN reaches the state-of-the-art results on 13 commonly employed datasets. To the best of our knowledge, it is the first experiment in GCANs applied to the problem of camera pose regression.

****

Contextual Affinity Distillation for Image Anomaly Detection

Jie Zhang, Masanori Suganuma, Takayuki Okatani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 149-158

Previous studies on unsupervised industrial anomaly detection mainly focus on 'structural' types of anomalies such as cracks and color contamination by matching or learning local feature representations. While achieving significantly high detection performance on this kind of anomaly, they are faced with 'logical' types of anomalies that violate the long-range dependencies such as a normal object placed in the wrong position. Noting the reverse distillation approaches that are under the encoder-decoder paradigm could learn from the high abstract level knowledge, we propose to use two students (local and global) to better mimic the teacher's local and global behavior in reverse distillation. The local student, which is used in previous studies mainly focuses on accurate local feature learning while the global student pays attention to learning global correlations. To further encourage the global student's learning to capture long-range dependencies, we design the global context condensing block (GCCB) and propose a contextual affinity loss for the student training and anomaly scoring. Experimental results show that the proposed method sets a new state-of-the-art performance on the MVTec LOCO AD dataset without using complex training techniques.

****

Semantic-Aware Video Representation for Few-Shot Action Recognition

Yutao Tang, Benjamín Béjar, René Vidal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6458-6468

Recent work on action recognition leverages 3D features and textual information to achieve state-of-the-art performance. However, most of the current few-shot action recognition methods still rely on 2D frame-level representations, often require additional components to model temporal relations, and employ complex distance functions to achieve accurate alignment of these representations. In addition, existing methods struggle to effectively integrate textual semantics, some resorting to concatenation or addition of textual and visual features, and some using text merely as an additional supervision without truly achieving feature fusion and information transfer from different modalities. In this work, we propose a simple yet effective Semantic-Aware Few-Shot Action Recognition (SAFSAR) model to address these issues. We show that directly leveraging a 3D feature extractor combined with an effective feature-fusion scheme, and a simple cosine similarity for classification can yield better performance without the need of extra components for temporal modeling or complex distance functions. We introduce an innovative scheme to encode the textual semantics into the video representation which adaptively fuses features from text and video, and encourages the visual encoder to extract more semantically consistent features. In this scheme, SAFSAR achieves alignment and fusion in a compact way. Experiments on five challenging few-shot action recognition benchmarks under various settings demonstrate that the proposed SAFSAR model significantly improves the state-of-the-art performance.

****

Adaptive Deep Neural Network Inference Optimization With EENet

Fatih Ilhan, Ka-Ho Chow, Sihao Hu, Tiansheng Huang, Selim Tekin, Wenqi Wei, Yanzhao Wu, Myungjin Lee, Ramana Kompella, Hugo Latapie, Gaowen Liu, Ling Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1373-1382

Well-trained deep neural networks (DNNs) treat all test samples equally during prediction. Adaptive DNN inference with early exiting leverages the observation that some test examples can be easier to predict than others. This paper presents EENet, a novel early-exiting scheduling framework for multi-exit DNN models. In stead of having every sample go through all DNN layers during prediction, EENet learns an early exit scheduler, which can intelligently terminate the inference earlier for certain predictions, which the model has high confidence of early exit. As opposed to previous early-exiting solutions with heuristics-based methods, our EENet framework optimizes an early-exiting policy to maximize model accuracy while satisfying the given per-sample average inference budget. Extensive experiments are conducted on four computer vision datasets (CIFAR-10, CIFAR-100, ImageNet, Cityscapes) and two NLP datasets (SST-2, AgNews). The results demonstrate that the adaptive inference by EENet can outperform the representative existing early exit techniques. We also perform a detailed visualization analysis of the comparison results to interpret the benefits of EENet.

****

MIVC: Multiple Instance Visual Component for Visual-Language Models

Wenyi Wu, Qi Li, Wenliang Zhong, Junzhou Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8117-8126

Vision-language models have been widely explored across a wide range of tasks and achieve satisfactory performance. However, it's under-explored how to consolidate entity understanding through a varying number of images and to align it with the pre-trained language models for generative tasks. In this paper, we propose MIVC, a general multiple instance visual component to bridge the gap between various image inputs with off-the-shelf vision-language models by aggregating visual representations in a permutation-invariant fashion through a neural network. We show that MIVC could be plugged into the visual-language models to improve the model performance consistently on visual question answering, classification and captioning tasks on a public available e-commerce dataset with multiple images per product. Furthermore, we show that the component provides insight into the contribution of each image to the downstream tasks.

****

Attentive Prototypes for Source-Free Unsupervised Domain Adaptive 3D Object Detection

Deepti Hegde, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3066-3076

3D object detection networks tend to be biased towards the data they are trained on. Evaluation on datasets captured in different locations, conditions or sensors than that of the training (source) data results in a drop in model performance due to the gap in distribution with the test (or target) data. Current methods for domain adaptation either assume access to source data during training, which may not be available due to privacy or memory concerns, or require a sequence of lidar frames as an input. We propose a single-frame approach for source-free, unsupervised domain adaptation of lidar-based 3D object detectors that uses class prototypes to mitigate the effect pseudo-label noise. Addressing the limitations of traditional feature aggregation methods for prototype computation in the presence of noisy labels, we utilize a transformer module to identify outlier ROI's that correspond to incorrect, over-confident annotations, and compute an attentive class prototype. Under an iterative training strategy, the losses associated with noisy pseudo labels are down-weighed and thus refined in the process of self-training. To validate the effectiveness of our proposed approach, we examine the domain shift associated with networks trained on large, label-rich datasets (such as the Waymo Open Dataset and nuScenes) and evaluate on smaller, label-poor datasets (such as KITTI) and vice-versa. We demonstrate our approach on two recent object detectors and achieve results that out-perform the other domain adaptation works.

****

Exploring the Impact of Rendering Method and Motion Quality on Model Performance When Using Multi-View Synthetic Data for Action Recognition

Stanislav Panev, Emily Kim, Sai Abhishek Si Namburu, Desislava Nikolova, Celso de Melo, Fernando De la Torre, Jessica Hodgins; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4592-4602

This paper explores the use of synthetic data in a human action recognition (HAR) task to avoid the challenges of obtaining and labeling real-world datasets. We introduce a new dataset suite comprising five datasets, eleven common human activities, three synchronized camera views (aerial and ground) in three outdoor environments, and three visual domains (real and two synthetic). For the synthetic data, two rendering methods (standard computer graphics and neural rendering) and two sources of human motions (motion capture and video-based motion reconstruction) were employed. We evaluated each dataset type by training popular activity recognition models and comparing the performance on the real test data. Our results show that synthetic data achieve slightly lower accuracy (4-8%) than real data. On the other hand, a model pre-trained on synthetic data and fine-tuned on limited real data surpasses the performance of either domain alone. Standard computer graphics (CG)-rendered data delivers better performance than the data generated from the neural-based rendering method. The results suggest that the quality of the human motions in the training data also affects the test results: motion capture delivers higher test accuracy. Additionally, a model trained on CG aerial view synthetic data exhibits greater robustness against camera viewpoint changes than one trained on real data. See the project page: http://humansensinglab.github.io/REMAG/.

****

ATS: Adaptive Temperature Scaling for Enhancing Out-of-Distribution Detection Methods

Gerhard Krumpl, Henning Avenhaus, Horst Possegger, Horst Bischof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3864-3873

Out-of-distribution (OOD) detection is essential to ensure the reliability and robustness of machine learning models in real-world applications. Post-hoc OOD detection methods have gained significant attention due to the fact that they offer the advantage of not requiring additional re-training, which could degrade model performance and increase training time. However, most existing post-hoc methods rely only on the encoder output (features), logits, or the softmax probability, meaning they have no access to information that might be lost in the feature extraction process. In this work, we address this limitation by introducing Adaptive Temperature Scaling (ATS), a novel approach that dynamically calculates a temperature value based on activations of the intermediate layers. Fusing this sample-specific adjustment with class-dependent logits, our ATS captures additional statistical information before they are lost in the feature extraction process, leading to a more robust and powerful OOD detection method. We conduct extensive experiments to demonstrate the efficacy of our approach. Notably, our method can be seamlessly combined with SOTA post-hoc OOD detection methods that rely on the logits, thereby enhancing their performance and improving their robustness.

****

Exploring Adversarial Robustness of Vision Transformers in the Spectral Perspective

Gihyun Kim, Juyeop Kim, Jong-Seok Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3976-3985

The Vision Transformer has emerged as a powerful tool for image classification tasks, surpassing the performance of convolutional neural networks (CNNs). Recently, many researchers have attempted to understand the robustness of Transformers against adversarial attacks. However, previous researches have focused solely on perturbations in the spatial domain. This paper proposes an additional perspective that explores the adversarial robustness of Transformers against frequency-selective perturbations in the spectral domain. To facilitate comparison between these two domains, an attack framework is formulated as a flexible tool for implementing attacks on images in both the spatial and spectral domains. The experiments reveal that Transformers rely more on phase and low frequency information, which can render them more vulnerable to frequency-selective attacks than CNNs. This work offers new insights into the properties and adversarial robustness of Transformers.

****

MotionAGFormer: Enhancing 3D Human Pose Estimation With a Transformer-GCNFormer Network

Soroush Mehraban, Vida Adeli, Babak Taati; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6920-6930

Recent transformer-based approaches have demonstrated excellent performance in 3D human pose estimation. However, they have a holistic view and by encoding global relationships between all the joints, they do not capture the local dependencies precisely. In this paper, we present a novel Attention-GCNFormer (AGFormer) block that divides the number of channels by using two parallel transformer and GCNFormer streams. Our proposed GCNFormer module exploits the local relationship between adjacent joints, outputting a new representation that is complementary to the transformer output. By fusing these two representation in an adaptive way, AGFormer exhibits the ability to better learn the underlying 3D structure. By stacking multiple AGFormer blocks, we propose MotionAGFormer in four different variants, which can be chosen based on the speed-accuracy trade-off. We evaluate our model on two popular benchmark datasets: Human3.6M and MPI-INF-3DHP. MotionAGFormer-B achieves state-of-the-art results, with P1 errors of 38.4 mm and 16.2 mm, respectively. Remarkably, it uses a quarter of the parameters and is three times more computationally efficient than the previous leading model on Human3.6M dataset. Code and models are available at https://github.com/TaatiTeam/MotionAGFormer.

****

Density-Based Flow Mask Integration via Deformable Convolution for Video People Flux Estimation

Chang-Lin Wan, Feng-Kai Huang, Hong-Han Shuai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6573-6582

Crowd counting is currently applied in many areas, such as transportation hubs and streets. However, most of the research still focuses on counting the number of people in a single image, and there is little research on solving the problem of calculating the number of non-repeated people in a video segment. Currently, multiple object tracking is mainly relied upon for video counting, but this method is not suitable for situations where the crowd density is too high. Therefore, we propose a Flow Mask Integration Deformable Convolution network (FMDC) combined with Intra-Frame Head Contrastive Learning (IFHC) to predict the situation of people entering and exiting the screen in a density-based manner. We verify that our proposed method is highly effective in densely populated situations and diverse scenes, and the experimental results show that our proposed method surpasses existing methods.

****

Learning Class and Domain Augmentations for Single-Source Open-Domain Generalization

Prathmesh Bele, Valay Bundele, Avigyan Bhattacharya, Ankit Jha, Gemma Roig, Biplab Banerjee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1816-1826

Single-source open-domain generalization (SS-ODG) addresses the challenge of labeled source domains with supervision during training and unlabeled novel target domains during testing. The target domain includes both known classes from the source domain and samples from previously unseen classes. Existing techniques for SS-ODG primarily focus on calibrating source-domain classifiers to identify open samples in the target domain. However, these methods struggle with visually fine-grained open-closed data, often misclassifying open samples as closed-set classes. Moreover, relying solely on a single source domain restricts the model's ability to generalize. To overcome these limitations, we propose a novel framework called SODG-NET that simultaneously synthesizes novel domains and generates pseudo-open samples using a learning-based objective, in contrast to the ad-hoc mixing strategies commonly found in the literature. Our approach enhances generalization by diversifying the styles of known class samples using a novel metric criterion and generates diverse pseudo-open samples to train a unified and confident multiclass classifier capable of handling both open and closed-set data. Extensive experimental evaluations conducted on multiple benchmarks consistently demonstrate the superior performance of SODG-NET compared to the literature.

****

RankDVQA: Deep VQA Based on Ranking-Inspired Hybrid Training
Chen Feng, Duolikun Danier, Fan Zhang, David Bull; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1648-1658

In recent years, deep learning techniques have shown significant potential for improving video quality assessment (VQA), achieving higher correlation with subjective opinions compared to conventional approaches. However, the development of deep VQA methods has been constrained by the limited availability of large-scale training databases and ineffective training methodologies. As a result, it is difficult for deep VQA approaches to achieve consistently superior performance and model generalization. In this context, this paper proposes new VQA methods based on a two-stage training methodology which motivates us to develop a large-scale VQA training database without employing human subjects to provide ground truth labels. This method was used to train a new transformer-based network architecture, exploiting quality ranking of different distorted sequences rather than minimizing the difference from the ground-truth quality labels. The resulting deep VQA methods (for both full reference and no reference scenarios), FR- and NR-RankDVQA, exhibit consistently higher correlation with perceptual quality compared to the state-of-the-art conventional and deep VQA methods, with average SROCC values of 0.8972 (FR) and 0.7791 (NR) over eight test sets without performing cross-validation. The source code of the proposed quality metrics and the large training database are available at https://chenfeng-bristol.github.io/RankDVQA.

****

Salient Object Detection for Images Taken by People With Vision Impairments
Jarek Reynolds, Chandra Kanth Nagesh, Danna Gurari; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8522-8531

Salient object detection is the task of producing a binary mask for an image that deciphers which pixels belong to the foreground object versus background. We introduce a new salient object detection dataset using images taken by people who are visually impaired who were seeking to better understand their surroundings, which we call VizWiz-SalientObject. Compared to seven existing datasets, VizWiz-SalientObject is the largest (i.e., 32,000 human-annotated images) and contains unique characteristics including a higher prevalence of text in the salient objects (i.e., in 68% of images) and salient objects that occupy a larger ratio of the images (i.e., on average, 50% coverage). We benchmarked ten modern models on our dataset. While most methods fall below human performance, struggling most for images with salient objects that are large, have less complex boundaries, and lack text as well as for lower quality images, one method one method is very close. To facilitate future extensions of this work, we publicly share the dataset at https://vizwiz.org/tasks-and-datasets/salient-object-detection.

****

HD-Fusion: Detailed Text-to-3D Generation Leveraging Multiple Noise Estimation
Jinbo Wu, Xiaobo Gao, Xing Liu, Zhengyang Shen, Chen Zhao, Haocheng Feng, Jingtuo Liu, Errui Ding; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3202-3211

In this paper, we study Text-to-3D content generation leveraging 2D diffusion priors to enhance the quality and detail of the generated 3D models. Recent progresses in text-to-3D have shown that employing high-resolution (e.g., 512 x 512) renderings can lead to the production of high-quality 3D models using latent diffusion priors. To enable rendering at even higher resolutions, which has the poten tial to further augment the quality and detail of the models, we propose a novel approach that combines multiple noise estimation processes with a pretrained diffusion prior. Distinct from the Bar-Tal et al.s' study which binds multiple denoised results [1] to generate images from texts, our approach integrates the computation of scoring distillation losses such as SDS loss and VSD loss which are essential techniques for the 3D content generation with 2D diffusion priors. We experimentally evaluated the proposed approach on XXX. The results show that the proposed approach can generate high-quality details more than the baselines.

****

pSTarC: Pseudo Source Guided Target Clustering for Fully Test-Time Adaptation
Manogna Sreenivas, Goirik Chakrabarty, Soma Biswas; Proceedings of the IEEE/CVF

Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2702-2710

Test Time Adaptation (TTA) is a pivotal concept in machine learning, enabling models to perform well in real-world scenarios, where test data distribution differs from training. In this work, we propose a novel approach called pseudo Source guided Target Clustering (pSTarC) addressing the relatively unexplored area of TTA under real-world domain shifts. This method draws inspiration from target clustering techniques and exploits the source classifier for generating pseudo source samples. The test samples are strategically aligned with these pseudo source samples, facilitating their clustering and thereby enhancing TTA performance. pSTarC operates solely within the fully test-time adaptation protocol, removing the need for actual source data. Experimental validation on a variety of domain shift datasets, namely VisDA, Office-Home, DomainNet-126, CIFAR-100C verifies pSTarC's effectiveness. This method exhibits significant improvements in prediction accuracy along with efficient computational requirements. Furthermore, we also demonstrate the universality of the pSTarC framework by showing its effectiveness for the continuous TTA framework.

****

FocusTune: Tuning Visual Localization Through Focus-Guided Sampling

Son Tung Nguyen, Alejandro Fontan, Michael Milford, Tobias Fischer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3606-3615

We propose FocusTune, a focus-guided sampling technique to improve the performance of visual localization algorithms. FocusTune directs a scene coordinate regression model towards regions critical for 3D point triangulation by exploiting key geometric constraints. Specifically, rather than uniformly sampling points across the image for training the scene coordinate regression model, we instead re-project 3D scene coordinates onto the 2D image plane and sample within a local neighborhood of the re-projected points. While our proposed sampling strategy is generally applicable, we showcase FocusTune by integrating it with the recently introduced Accelerated Coordinate Encoding (ACE) model. Our results demonstrate that FocusTune both improves or matches state-of-the-art performance whilst keeping ACE's appealing low storage and compute requirements, for example reducing translation error from 25 to 19 and 17 to 15 cm for single and ensemble models, respectively, on the Cambridge Landmarks dataset. This combination of high performance and low compute and storage requirements is particularly promising for applications in areas like mobile robotics and augmented reality. We made our code available at https://github.com/sontung/focus-tune.

****

Improving Normalization With the James-Stein Estimator

Seyedalireza Khoshsirat, Chandra Kambhamettu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2041-2051

Stein's paradox holds considerable sway in high-dimensional statistics, highlighting that the sample mean, traditionally considered the de facto estimator, might not be the most efficacious in higher dimensions. To address this, the James-Stein estimator proposes an enhancement by steering the sample means toward a more centralized mean vector. In this paper, first, we establish that normalization layers in deep learning use inadmissible estimators for mean and variance. Next, we introduce a novel method to employ the James-Stein estimator to improve the estimation of mean and variance within normalization layers. We evaluate our method on different computer vision tasks: image classification, semantic segmentation, and 3D object classification. Through these evaluations, it is evident that our improved normalization layers consistently yield superior accuracy across all tasks without extra computational burden. Moreover, recognizing that a plethora of shrinkage estimators surpass the traditional estimator in performance, we study two other prominent shrinkage estimators: Ridge and LASSO. Additionally, we provide visual representations to intuitively demonstrate the impact of shrinkage on the estimated layer statistics. Finally, we study the effect of regularization and batch size on our modified batch normalization. The studies show that our method is less sensitive to batch size and regularization, improving accuracy under various setups.

****

Depth From Asymmetric Frame-Event Stereo: A Divide-and-Conquer Approach

Xihao Chen, Wenming Weng, Yueyi Zhang, Zhiwei Xiong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3045-3054

Event cameras asynchronously measure brightness changes in a scene without motion blur or saturation, while frame cameras capture images with dense intensity and fine details at a fixed rate. The exclusive advantages of the two modalities make depth estimation from Stereo Asymmetric Frame-Event (SAFE) systems appealing. However, due to the inevitable information absence of one modality in certain challenging regions, existing stereo matching methods lose efficacy for asymmetric inputs from SAFE systems. In this paper, we propose a divide-and-conquer approach that decomposes depth estimation from SAFE systems into three sub-tasks, i.e., frame-event stereo matching, frame-based Structure-from-Motion (SfM), and event-based SfM. In this way, the above challenging regions are addressed by monocular SfM, which estimates robust depth with two views belonging to the same functioning modality. Moreover, we propose a dual sampling strategy to construct cost volumes with identical spatial locations and depth hypotheses for different sub-tasks, which enables sub-task fusion at the cost volume level. To tackle the occlusion issue raised by the sampling strategy, we further introduce a temporal fusion scheme to utilize long-term sequential inputs with multi-view information. Experimental results validate the superior performance of our method over existing solutions.

****

Framework-Agnostic Semantically-Aware Global Reasoning for Segmentation

Mir Rayat Imtiaz Hossain, Leonid Sigal, James J. Little; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 999-1009

Recent advances in pixel-level tasks (e.g. segmentation) illustrate the benefit of of long-range interactions between aggregated region-based representations that can enhance local features. However, such aggregated representations, often in the form of attention, fail to model the underlying semantics of the scene (e.g. individual objects and, by extension, their interactions). In this work, we address the issue by proposing a component that learns to project image features into latent representations and reason between them using a transformer encoder to generate contextualized and scene-consistent representations which are fused with original image features. Our design encourages the latent regions to represent semantic concepts by ensuring that the activated regions are spatially disjoint and the union of such regions corresponds to a connected object segment. The proposed semantic global reasoning (SGR) component is end-to-end trainable and can be easily added to a wide variety of backbones (CNN or transformer-based) and segmentation heads (per-pixel or mask classification) to consistently improve the segmentation results on different datasets. In addition, our latent tokens are semantically interpretable and diverse and provide a rich set of features that can be transferred to downstream tasks like object detection and segmentation, with improved performance. Furthermore, we also proposed metrics to quantify the semantics of latent tokens at both class & instance level.

****

Self-Supervised Relation Alignment for Scene Graph Generation

Bicheng Xu, Renjie Liao, Leonid Sigal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1339-1349

The goal of scene graph generation is to predict a graph from an input image, where nodes correspond to identified and localized objects and edges to their corresponding interaction predicates. Existing methods are trained in a fully supervised manner and focus on message passing mechanisms, loss functions, and/or bias mitigation. In this work we introduce a simple-yet-effective self-supervised relational alignment regularization designed to improve the scene graph generation performance. The proposed alignment is general and can be combined with any existing scene graph generation framework, where it is trained alongside the original model's objective. The alignment is achieved through distillation, where an a

uxiliary relation prediction branch, that mirrors and shares parameters with the supervised counterpart, is designed. In the auxiliary branch, relational input features are partially masked prior to message passing and predicate prediction. The predictions for masked relations are then aligned with the supervised counterparts after the message passing. We illustrate the effectiveness of this self-supervised relational alignment in conjunction with two scene graph generation architectures, SGTR and Neural Motifs, and show that in both cases we achieve significantly improved performance.
****

Tackling Data Bias in MUSIC-AVQA: Crafting a Balanced Dataset for Unbiased Question-Answering

Xiulong Liu, Zhikang Dong, Peng Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4478-4487

In recent years, there has been a growing emphasis on the intersection of audio, vision, and text modalities, driving forward the advancements in multimodal research. However, strong bias that exists in any modality can lead to the model neglecting the others. Consequently, the model's ability to effectively reason across these diverse modalities is compromised, impeding further advancement. In recent years, there has been a growing emphasis on the intersection of audio, vision, and text modalities, driving forward the advancements in multimodal research. However, strong bias that exists in any modality can lead to the model neglecting the others. Consequently, the model's ability to effectively reason across these diverse modalities is compromised, impeding further advancement. In this paper, we meticulously review each question type from the original dataset, selecting those with pronounced answer biases. To counter these biases, we gather complementary videos and questions, ensuring that no answers have outstanding skewed distribution. In particular, for binary questions, we strive to ensure that both answers are almost uniformly spread within each question category. As a result, we construct a new dataset, named MUSIC-AVQA v2.0, which is more challenging and we believe could better foster the progress of AVQA task. Furthermore, we present a novel baseline model that delves deeper into the audio-visual-text interrelation. On MUSIC-AVQA v2.0, this model surpasses all the existing benchmarks, improving accuracy by 2% on MUSIC-AVQA v2.0, setting a new state-of-the-art performance.
****

RPCANet: Deep Unfolding RPCA Based Infrared Small Target Detection

Fengyi Wu, Tianfang Zhang, Lei Li, Yian Huang, Zhenming Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4809-4818

Deep learning (DL) networks have achieved remarkable performance in infrared small target detection (ISTD). However, these structures exhibit a deficiency in interpretability and are widely regarded as black boxes, as they disregard domain knowledge in ISTD. To alleviate this issue, this work proposes an interpretable deep network for detecting infrared dim targets, dubbed RPCANet. Specifically, our approach formulates the ISTD task as sparse target extraction, low-rank background estimation, and image reconstruction in a relaxed Robust Principle Component Analysis (RPCA) model. By unfolding the iterative optimization updating steps into a deep-learning framework, time-consuming and complex matrix calculations are replaced by theory-guided neural networks. RPCANet detects targets with clear interpretability and preserves the intrinsic image feature, instead of directly transforming the detection task into a matrix decomposition problem. Extensive experiments substantiate the effectiveness of our deep unfolding framework and demonstrate its trustworthy results, surpassing baseline methods in both qualitative and quantitative evaluations.
****

GLAD: Global-Local View Alignment and Background Debiasing for Unsupervised Video Domain Adaptation With Large Domain Gap

Hyogun Lee, Kyungho Bae, Seong Jong Ha, Yumin Ko, Gyeong-Moon Park, Jinwoo Choi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6816-6825

In this work, we tackle the challenging problem of unsupervised video domain adaptation (UVDA) for action recognition. We specifically focus on scenarios with a substantial domain gap, in contrast to existing works primarily deal with small domain gaps between labeled source domains and unlabeled target domains. To establish a more realistic setting, we introduce a novel UVDA scenario, denoted as Kinetics->BABEL, with a more considerable domain gap in terms of both temporal dynamics and background shifts. To tackle the temporal shift, i.e., action duration difference between the source and target domains, we propose a global-local view alignment approach. To mitigate the background shift, we propose to learn temporal order sensitive representations by temporal order learning and background invariant representations by background augmentation. We empirically validate that the proposed method shows significant improvement over the existing methods on the Kinetics->BABEL dataset with a large domain gap.

****

Learning To Compose SuperWeights for Neural Parameter Allocation Search

Piotr Teterwak, Soren Nelson, Nikoli Dryden, Dina Bashkirova, Kate Saenko, Bryan A. Plummer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2751-2760

Neural parameter allocation search (NPAS) automates parameter sharing by obtaining weights for a network given an arbitrary, fixed parameter budget. Prior work has two major drawbacks we aim to address. First, there is a dis- connect in the sharing pattern between the search and train- ing steps, where weights are warped for layers of different sizes during the search to measure similarity, but not during training, resulting in reduced performance. To address this, we generate layer weights by learning to compose sets of SuperWeights, which represent a group of trainable parameters. These SuperWeights are created to be large enough so they can be used to represent any layer in the network, but small enough that they are computationally efficient. The second drawback we address is the method of measuring similarity between shared parameters. Whereas prior work compared the weights themselves, we argue this does not take into account the amount of conflict between the shared weights. Instead, we use gradient information to identify layers with shared weights that wish to diverge from each other. We demonstrate that our SuperWeight Networks consistently boost performance over the state-of-the-art on the ImageNet and CIFAR datasets in the NPAS setting. We further show that our approach can generate parameters for many network architectures using the same set of weights. This enables us to support tasks like efficient ensembling and anytime prediction, outperforming fully-parameterized ensembles with 17% fewer parameters.

****

Second-Order Graph ODEs for Multi-Agent Trajectory Forecasting

Song Wen, Hao Wang, Di Liu, Qilong Zhangli, Dimitris Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5101-5110

Trajectory forecasting of multiple agents is a fundamental task that has applications in various fields, such as autonomous driving, physical system modeling and smart cities. It is challenging because agent interactions and underlying continuous dynamics jointly affect its behavior. Existing approaches often rely on Graph Neural Networks (GNNs) or Transformers to extract agent interaction features. However, they tend to neglect how the distance and velocity information between agents impact their interactions dynamically. Moreover, previous methods use RNNs or first-order Ordinary Differential Equations (ODEs) to model temporal dynamics, which may lack interpretability with respect to how each agent is driven by interactions. To address these challenges, this paper proposes the Agent Graph ODE, a novel approach that models agent interactions and continuous second-order dynamics explicitly. Our method utilizes a variational autoencoder architecture, incorporating spatial-temporal Transformers with distance information and dynamic interaction graph construction in the encoder module. In the decoder module, we employ GNNs with distance information to model agent interactions, and use coupled second-order ODEs to capture the underlying continuous dynamics by modeling the relationship between acceleration and agent interactions. Experimental results show that our proposed Agent Graph ODE outperforms state-of-the-art methods in prediction accuracy. Moreover, our method performs well in sudden situations not seen in the training dataset.

****

FOUND: Foot Optimization With Uncertain Normals for Surface Deformation Using Synthetic Data

Oliver Boyne, Gwangbin Bae, James Charles, Roberto Cipolla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8097-8106

Surface reconstruction from multi-view images is a challenging task, with solutions often requiring a large number of sampled images with high overlap. We seek to develop a method for few-view reconstruction, for the case of the human foot. To solve this task, we must extract rich geometric cues from RGB images, before carefully fusing them into a final 3D object. Our FOUND approach tackles this, with 4 main contributions: (i) SynFoot, a synthetic dataset of 50,000 photorealistic foot images, paired with ground truth surface normals and keypoints; (ii) an uncertainty-aware surface normal predictor trained on our synthetic dataset; (iii) an optimization scheme for fitting a generative foot model to a series of images; and (iv) a benchmark dataset of calibrated images and high resolution ground truth geometry. We show that our normal predictor outperforms all off-the-shelf equivalents significantly on real images, and our optimization scheme outperforms state-of-the-art photogrammetry pipelines, especially for a few-view setting. We release our synthetic dataset and baseline 3D scans to the research community.

****

Unsupervised Event-Based Video Reconstruction

Gereon Fox, Xingang Pan, Ayush Tewari, Mohamed Elgharib, Christian Theobalt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4179-4188

Event cameras report events whenever an individual pixel changes brightness. The discrete and asynchronous nature of events makes recovering pixel brightness signals a challenging task, even if conventional brightness frames are recorded along with events. Recent works have addressed this task with neural networks, which tend to be biased towards their training distribution. All methods need to deal with noise in the events to produce very high output framerates. We introduce a new approach to event-based reconstruction, not learning-based: Our model assigns each event an explicit confidence weight to account for the uncertainty arising from noise. We also introduce a novel loss term to balance confidences against each other and show that interpolation of brightness signals between events can benefit from Bezier curves. We demonstrate that allowing brightness changes between exposures can improve reconstruction quality. Our evaluation shows that our method improves the state of the art in the tasks of event-based deblurring and event-based frame interpolation.

****

Can CLIP Help Sound Source Localization?

Sooyoung Park, Arda Senocak, Joon Son Chung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5711-5720

Large-scale pre-trained image-text models demonstrate remarkable versatility across diverse tasks, benefiting from their robust representational capabilities and effective multimodal alignment. We extend the application of these models, specifically CLIP, to the domain of sound source localization. Unlike conventional approaches, we employ the pre-trained CLIP model without explicit text input, relying solely on the audio-visual correspondence. To this end, we introduce a framework that translates audio signals into tokens compatible with CLIP's text encoder, yielding audio-driven embeddings. By directly using these embeddings, our method generates audio-grounded masks for the provided audio, extracts audio-grounded image features from the highlighted regions, and aligns them with the audio-driven embeddings using the audio-visual correspondence objective. Our findings suggest that utilizing pre-trained image-text models enable our model to generate more complete and compact localization maps for the sounding objects. Extens ive experiments show that our method outperforms state-of-the-art approaches by a significant margin.

****

FastSR-NeRF: Improving NeRF Efficiency on Consumer Devices With a Simple Super-Resolution Pipeline

Chien-Yu Lin, Qichen Fu, Thomas Merth, Karren Yang, Anurag Ranjan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6036-6045

Super-resolution (SR) techniques have recently been proposed to upscale the outputs of neural radiance fields (NeRF) and generate high-quality images with enhanced inference speeds. However, existing NeRF+SR methods increase training overhead by using extra input features, loss functions, or expensive training procedures such as knowledge distillation. In this paper, we aim to leverage SR for efficiency gains without costly training or architectural changes. Specifically, we build a simple NeRF+SR pipeline that directly combines existing modules, and we propose a lightweight augmentation technique, random patch sampling, for training. Compared to existing NeRF+SR methods, our pipeline mitigates the SR computing overhead and can be trained up to 23x faster, making it feasible to run on consumer devices such as the Apple MacBook. Experiments show that our pipeline can upscale NeRF outputs by 2-4x while maintaining high quality, increasing inference speeds by up to 18x on an NVIDIA V100 GPU and 12.8x on an M1 Pro chip. We conclude that SR can be a simple but effective technique for improving the efficiency of NeRF models for consumer devices.

****

Online Class-Incremental Learning for Real-World Food Image Classification

Siddeshwar Raghavan, Jiangpeng He, Fengqing Zhu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8195-8204

Food image classification is essential for monitoring health and tracking dietary in image-based dietary assessment methods. However, conventional systems often rely on static datasets with fixed classes and uniform distribution. In contrast, real-world food consumption patterns, shaped by cultural, economic, and personal influences, involve dynamic and evolving data. Thus, require the classification system to cope with continuously evolving data. Online Class Incremental Learning (OCIL) addresses the challenge of learning continuously from a single-pass data stream while adapting to the new knowledge and reducing catastrophic forgetting. Experience Replay (ER) based OCIL methods store a small portion of previous data and have shown encouraging performance. However, most existing OCIL works assume that the distribution of encountered data is perfectly balanced, which rarely happens in real-world scenarios. In this work, we explore OCIL for real-world food image classification by first introducing a probabilistic framework to simulate realistic food consumption scenarios. Subsequently, we present an attachable Dynamic Model Update (DMU) module designed for existing ER methods, which enables the selection of relevant images for model training, addressing challenges arising from data repetition and imbalanced sample occurrences inherent in realistic food consumption patterns within the OCIL framework. Our performance evaluation demonstrates significant enhancements compared to established ER methods, showing great potential for lifelong learning in real-world food image classification scenarios. The code of our method is publicly accessible at https://gitlab.com/viper-purdue/OCIL-real-world-food-image-classification

****

United We Stand, Divided We Fall: UnityGraph for Unsupervised Procedure Learning From Videos

Siddhant Bansal, Chetan Arora, C. V. Jawahar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6509-6519

Given multiple videos of the same task, procedure learning addresses identifying the key-steps and determining their order to perform the task. For this purpose, existing approaches use the signal generated from a pair of videos. This makes key-steps discovery challenging as the algorithms lack inter-videos perspective. Instead, we propose an unsupervised Graph-based Procedure Learning (GPL) framework. GPL consists of the novel UnityGraph that represents all the videos of a t ask as a graph to obtain both intra-video and inter-videos context. Further, to obtain similar embeddings for the same key-steps, the embeddings of UnityGraph are updated in an unsupervised manner using the Node2Vec algorithm. Finally, to identify the key-steps, we cluster the embeddings using KMeans. We test GPL on benchmark ProceL, CrossTask, and EgoProceL datasets and achieve an average improvement of 2% on third-person datasets and 3.6% on EgoProceL over the state-of-the-art.

****

3D Face Style Transfer With a Hybrid Solution of NeRF and Mesh Rasterization

Jianwei Feng, Prateek Singhal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3504-3513

Style transfer for human face has been widely researched in recent years. Majority of the existing approaches work in 2D image domain and have 3D inconsistency issue when applied on different viewpoints of the same face. In this paper, we tackle the problem of 3D face style transfer which aims at generating stylized novel views of a 3D human face with multi-view consistency. We propose to use a neural radiance field (NeRF) to represent 3D human face and combine it with 2D style transfer to stylize the 3D face. We find that directly training a NeRF on stylized images from 2D style transfer brings in 3D inconsistency issue and causes blurriness. On the other hand, training a NeRF jointly with 2D style transfer objectives shows poor convergence due to the identity and head pose gap between style image and content image. It also poses challenge in training time and memory due to the need of volume rendering for full image to apply style transfer loss functions. We therefore propose a hybrid framework of NeRF and mesh rasterization to combine the benefits of high fidelity geometry reconstruction of NeRF and fast rendering speed of mesh. Our framework consists of three stages: 1. Training a NeRF model on input face images to learn the 3D geometry; 2. Extracting a mesh from the trained NeRF model and optimizing it with style transfer objectives via differentiable rasterization; 3. Training a new color network in NeRF conditioned on a style embedding to enable arbitrary style transfer to the 3D face. Experiment results show that our approach generates high quality face style transfer with great 3D consistency, while also enabling a flexible style control.

****

USDN: A Unified Sample-Wise Dynamic Network With Mixed-Precision and Early-Exit

Ji-Ye Jeon, Xuan Truong Nguyen, Soojung Ryu, Hyuk-Jae Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 646-654

To reduce computation in deep neural network inference, a promising approach is to design a network with multiple internal classifiers (ICs) and adaptively select an execution path based on the complexity of a given input. However, quantizing an input-adaptive network, a must-do task for network deployment on edge devices, is a non-trivial task due to jointly allocating its computation budget along with network layers and IC locations. In this paper, we propose Unified Sample-wise Dynamic Network (USDN) with a mixed-precision and early-exit framework that obtains both the optimal location of ICs and layer-wise bit configurations under a given computation budget. The proposed USDN comprises multiple groups of layers, with each group representing a varying degree of complexity for input samples. Experimental results demonstrate that our approach reduces computational cost of the previous work by 12.78% while achieving higher accuracy on ImageNet dataset.

****

Learn To Unlearn for Deep Neural Networks: Minimizing Unlearning Interference With Gradient Projection

Tuan Hoang, Santu Rana, Sunil Gupta, Svetha Venkatesh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4819-4828

Recent data-privacy laws have sparked interest in machine unlearning, which involves removing the effect of specific training samples from a learnt model as if they were never present in the original training dataset. The challenge of machine unlearning is to discard information about the "forget" data in the learnt mo del without altering the knowledge about the remaining dataset and to do so more efficiently than the naive retraining approach. To achieve this, we adopt a projected-gradient based learning method, named as Projected-Gradient Unlearning (PGU), in which the model takes steps in the orthogonal direction to the gradient subspaces deemed unimportant for the retaining dataset, so as to its knowledge is preserved. By utilizing Stochastic Gradient Descent (SGD) to update the model weights, our method can efficiently scale to any model and dataset size. We provide empirically evidence to demonstrate that our unlearning method can produce models that behave similar to models retrained from scratch across various metrics even when the training dataset is no longer accessible. Our code is available at https://github.com/hnanhtuan/projected_gradient_unlearning.

****

Human Motion Aware Text-to-Video Generation With Explicit Camera Control

Taehoon Kim, ChanHee Kang, JaeHyuk Park, Daun Jeong, ChangHee Yang, Suk-Ju Kang, Kyeongbo Kong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5081-5090

With the rise in expectations related to generative models, text-to-video (T2V) models are being actively studied. Existing text-to-video models have limitations such as in generating complex movements replicating human motions. These model often generate unintended human motions, and the scale of the subject is incorrect. To overcome these limitations and generate high-quality videos that depict human motion under plausible viewing angles, we propose a two stage framework in this study. In the first stage a text-driven human motion generation network generates three-dimensional (3D) human motion from input text prompts and then motion-to-skeleton projection module projects generated motions onto a two-dimensional (2D) skeleton. In the second stage, the projected skeletons are used to generate a video in which the movements of a subject are well-represented. We demonstrated that the proposed framework quantitatively and qualitatively outperforms the existing T2V models. Previously reported human motion generation models use texts only or texts and human skeletons. However, our framework only uses texts and outputs a video related to human motion. Moreover, our framework benefits from using skeleton as an additional condition in the text-to-human motion generation networks. To the best of our knowledge, our framework is the first of its kind that uses text-driven human motion generation networks to generate high-quality videos related to human motions. The corresponding codes are available at https://github.com/CSJasper/HMTV.

****

Beyond SOT: Tracking Multiple Generic Objects at Once

Christoph Mayer, Martin Danelljan, Ming-Hsuan Yang, Vittorio Ferrari, Luc Van Gool, Alina Kuznetsova; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6826-6836

Generic Object Tracking (GOT) is the problem of tracking target objects, specified by bounding boxes in the first frame of a video. While the task has received much attention in the last decades, researchers have almost exclusively focused on the single object setting. However multi-object GOT poses its own challenges and is more attractive in real-world applications. We attribute the lack of research interest into this problem to the absence of suitable benchmarks. In this work, we introduce a new large-scale GOT benchmark, LaGOT, containing multiple annotated target objects per sequence. Our benchmark allows users to tackle key remaining challenges in GOT, aiming to increase robustness and reduce computation through joint tracking of multiple objects simultaneously. In addition, we propose a transformer-based GOT tracker baseline capable of joint processing of multiple objects through shared computation. Our approach achieves a 4x faster run-time in case of 10 concurrent objects compared to tracking each object independently and outperforms existing single object trackers on our new benchmark. In addition, our approach achieves highly competitive results on single-object GOT datasets, setting a new state of the art on TrackingNet with a success rate AUC of 84.4%. Our benchmark, code, and trained models will be made publicly available.

****

Revisiting Latent Space of GAN Inversion for Robust Real Image Editing

Kai Katsumata, Duc Minh Vo, Bei Liu, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5313-5322

We present a generative adversarial network (GAN) inversion with high reconstruction and editing quality. GAN inversion algorithms with expressive latent spaces produce near-perfect inversion but are not robust to editing operations in latent space, leading to undesirable edited images, a phenomenon known as the trade-off between reconstruction and editing quality. To cope with the trade-off, we revisit the hyperspherical prior of StyleGANs Z and propose to combine an extended space of Z with highly capable inversion algorithms. Our approach maintains the reconstruction quality of seminal GAN inversion methods while improving their editing quality owing to the constrained nature of Z. Through comprehensive experiments with several GAN inversion algorithms, we demonstrate that our approach enhances image editing quality in 2D/3D GANs.

****

Robust TRISO-Fueled Pebble Identification by Digit Recognition

Roshan Kenia, Jihane Mendil, Ahmed Jasim, Muthanna Al-Dahhan, Zhaozheng Yin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8157-8165

Nuclear power plays a vital role in providing reliable and clean energy to fulfill increasing demands in electricity worldwide. It continues to be an essential source of national power supply as growing concerns about fossil fuel depletion, global warming, and emissions require utilizing sustainable energy sources. One area contributing to the growth of nuclear power is the development of reactors that have enhanced protection and security, thermal efficiency, and design. Reactor efficiency can be studied by the burnup that occurs when a TRISO-fueled pebble is inserted into the nuclear core and subsequently removed. The levels of burnup are measured based on the length of time the pebble spends within the core. In our design, each pebble is numbered by multiple digits printed in six locations using Ultra-High Temperature Ceramic paint. Naturally, computer vision techniques can be used to identify and time each pebble based on its digits as it enters and exits the core. We present a deep learning approach that successfully tags each pebble by identifying its digits from a video stream of the entrance and exit of the core. In a multi-step method, we extract only the clearest and most useful views of the pebble's digits to classify as it rolls by. This algorithm is robust against issues that occur for objects in movement such as motion blur, rotations, and glare. We outperform other state-of-the-art optical character recognition (OCR) models that fail to identify digits that are in motion. Our approach creates a safer and more efficient way to measure burnup within a core while contributing to the improvement of nuclear power produced by reactors.

****

Evidential Uncertainty Quantification: A Variance-Based Perspective

Ruxiao Duan, Brian Caffo, Harrison X. Bai, Haris I. Sair, Craig Jones; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2132-2141

Uncertainty quantification of deep neural networks has become an active field of research and plays a crucial role in various downstream tasks such as active learning. Recent advances in evidential deep learning shed light on the direct quantification of aleatoric and epistemic uncertainties with a single forward pass of the model. Most traditional approaches adopt an entropy-based method to derive evidential uncertainty in classification, quantifying uncertainty at the sample level. However, the variance-based method that has been widely applied in regression problems is seldom used in the classification setting. In this work, we adapt the variance-based approach from regression to classification, quantifying classification uncertainty at the class level. The variance decomposition technique in regression is extended to class covariance decomposition in classification based on the law of total covariance, and the class correlation is also derived from the covariance. Experiments on cross-domain datasets are conducted to illustrate that the variance-based approach not only results in similar accuracy as the entropy-based one in active domain adaptation but also brings information a

bout class-wise uncertainties as well as between-class correlations. The code is available at https://github.com/KerryDRX/EvidentialADA. This alternative means of evidential uncertainty quantification will give researchers more options when class uncertainties and correlations are important in their applications.
****

ICF-SRSR: Invertible Scale-Conditional Function for Self-Supervised Real-World Single Image Super-Resolution

Reyhaneh Neshatavar, Mohsen Yavartanoo, Sanghyun Son, Kyoung Mu Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1557-1567

Single image super-resolution (SISR) is a challenging ill-posed problem that aims to up-sample a given low-resolution (LR) image to a high-resolution (HR) counterpart. Due to the difficulty in obtaining real LR-HR training pairs, recent approaches are trained on simulated LR images degraded by simplified down-sampling operators, e.g., bicubic. Such an approach can be problematic in practice due to the large gap between the synthesized and real-world LR images. To alleviate the issue, we propose a novel Invertible scale-Conditional Function (ICF), which can scale an input image and then restore the original input with different scale conditions. Using the proposed ICF, we construct a novel self-supervised SISR framework (ICF-SRSR) to handle the real-world SR task without using any paired/unpaired training data. Furthermore, our ICF-SRSR can generate realistic and feasible LR-HR pairs, which can make existing supervised SISR networks more robust. Extensive experiments demonstrate the effectiveness of our method in handling SISR in a fully self-supervised manner. Our ICF-SRSR demonstrates superior performance compared to the existing methods trained on synthetic paired images in real-world scenarios and exhibits comparable performance compared to state-of-the-art supervised/unsupervised methods on public benchmark datasets. The code is available from this link.
****

PATROL: Privacy-Oriented Pruning for Collaborative Inference Against Model Inversion Attacks

Shiwei Ding, Lan Zhang, Miao Pan, Xiaoyong Yuan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4716-4725

Collaborative inference has been a promising solution to enable resource-constrained edge devices to perform inference using state-of-the-art deep neural networks (DNNs). In collaborative inference, the edge device first feeds the input to a partial DNN locally and then uploads the intermediate result to the cloud to complete the inference. However, recent research indicates model inversion attacks (MIAs) can reconstruct input data from intermediate results, posing serious privacy concerns for collaborative inference. Existing perturbation and cryptography techniques are inefficient and unreliable in defending against MIAs while performing accurate inference. This paper provides a viable solution, named PATROL, which develops privacy-oriented pruning to balance privacy, efficiency, and utility of collaborative inference. PATROL takes advantage of the fact that later layers in a DNN can extract more task-specific features. Given limited local resources for collaborative inference, PATROL intends to deploy more layers at the edge based on pruning techniques to enforce task-specific features for inference and reduce task-irrelevant but sensitive features for privacy preservation. To achieve privacy-oriented pruning, PATROL introduces two key components: Lipschitz regularization and adversarial reconstruction training, which increase the reconstruction errors by reducing the stability of MIAs and enhance the target inference model by adversarial training, respectively. On a real-world collaborative inference task, vehicle re-identification, we demonstrate the superior performance of PATROL in terms of against MIAs.
****

Collage Diffusion

Vishnu Sarukkai, Linden Li, Arden Ma, Christopher Ré, Kayvon Fatahalian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4208-4217

We seek to give users precise control over diffusion-based image generation by m odeling complex scenes as sequences of layers, which define the desired spatial arrangement and visual attributes of objects in the scene. Collage Diffusion harmonizes the input layers to make objects fit together---the key challenge involves minimizing changes in the positions and key visual attributes of the input layers while allowing other attributes to change in the harmonization process. We ensure that objects are generated in the correct locations by modifying text-image cross-attention with the layers' alpha masks. We preserve key visual attributes of input layers by learning specialized text representations per layer and by extending prior diffusion-based control mechanisms to operate on layers. Layer input allows users to control the extent of image harmonization on a per-object basis, and users can even iteratively edit individual objects in generated images while keeping other objects fixed. By leveraging the rich information present in layer input, Collage Diffusion generates globally harmonized images that maintain desired object characteristics better than prior approaches.

****

Camera-Independent Single Image Depth Estimation From Defocus Blur

Lahiru Wijayasingha, Homa Alemzadeh, John A. Stankovic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3749-3758

Monocular depth estimation is an important step in many downstream tasks in machine vision. We address the topic of estimating monocular depth from defocus blur which can yield more accurate results than the semantic based depth estimation methods. The existing monocular depth from defocus techniques are sensitive to the particular camera that the images are taken from. We show how several camera-related parameters affect the defocus blur using optical physics equations and how they make the defocus blur depend on these parameters. The simple correction procedure we propose can alleviate this problem which does not require any retraining of the original model. We created a synthetic dataset which can be used to test the camera independent performance of depth from defocus blur models. We evaluate our model on both synthetic and real datasets (DDFF12 and NYU depth V2) obtained with different cameras and show that our methods are significantly more robust to the changes of cameras.

****

Wakening Past Concepts Without Past Data: Class-Incremental Learning From Online Placebos

Yaoyao Liu, Yingying Li, Bernt Schiele, Qianru Sun; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2226-2235

Not forgetting old class knowledge is a key challenge for class-incremental learning (CIL) when the model continuously adapts to new classes. A common technique to address this is knowledge distillation (KD), which penalizes prediction inconsistencies between old and new models. Such prediction is made with almost new class data, as old class data is extremely scarce due to the strict memory limitation in CIL. In this paper, we take a deep dive into KD losses and find that "using new class data for KD" not only hinders the model adaption (for learning new classes) but also results in low efficiency for preserving old class knowledge. We address this by "using the placebos of old classes for KD", where the placebos are chosen from a free image stream, such as Google Images, in an automatical and economical fashion. To this end, we train an online placebo selection policy to quickly evaluate the quality of streaming images (good or bad placebos) and use only good ones for one-time feed-forward computation of KD. We formulate the policy training process as an online Markov Decision Process (MDP), and introduce an online learning algorithm to solve this MDP problem without causing much computation costs. In experiments, we show that our method 1) is surprisingly effective even when there is no class overlap between placebos and original old class data, 2) does not require any additional supervision or memory budget, and 3) significantly outperforms a number of top-performing CIL methods, in particular when using lower memory budgets for old class exemplars, e.g., five exemplars per class.

****

Fine-Grained Alignment for Cross-Modal Recipe Retrieval

Muntasir Wahed, Xiaona Zhou, Tianjiao Yu, Ismini Lourentzou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5584-5593

Vision-language pre-trained models have exhibited significant advancements in various multimodal and unimodal tasks in recent years, including cross-modal recipe retrieval. However, a persistent challenge in multimodal frameworks is the lack of alignment between the encoders of different modalities. Although previous works addressed image and recipe embedding alignment, the alignment of individual recipe components has been overlooked. To address this gap, we present Fine-grained Alignment for Recipe Embeddings (FARM), a cross-modal retrieval approach that aligns the encodings of recipe components, including titles, ingredients, and instructions, within a shared representation space alongside corresponding image embeddings. Moreover, we introduce a hyperbolic loss function to effectively capture the similarity information inherent in recipe classes. FARM improves Recall@1 by 1.4% for image-to-recipe and 1.0 for recipe-to-image retrieval. Additionally, FARM achieves up to 6.1% and 15.1% performance improvement in image-to-recipe retrieval tasks, when just one and two components of the recipe are available, respectively. Comprehensive qualitative analysis of retrieved images for various recipes showcases the semantic capabilities of our trained models. Code is available at https://github.com/PLAN-Lab/FARM.

****

NOMAD: A Natural, Occluded, Multi-Scale Aerial Dataset, for Emergency Response Scenarios

Arturo Miguel Russell Bernal, Walter Scheirer, Jane Cleland-Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8584-8595

With the increasing reliance on small Unmanned Aerial Systems (sUAS) for Emergency Response Scenarios, such as Search and Rescue, the integration of computer vision capabilities has become a key factor in mission success. Nevertheless, computer vision performance for detecting humans severely degrades when shifting from ground to aerial views. Several aerial datasets have been created to mitigate this problem, however, none of them has specifically addressed the issue of occlusion, a critical component in Emergency Response Scenarios. Natural Occluded Multi-scale Aerial Dataset (NOMAD) presents a benchmark for human detection under occluded aerial views, with five different aerial distances and rich imagery variance. NOMAD is composed of 100 different Actors, all performing sequences of walking, laying and hiding. It includes 42,825 frames, extracted from 5.4k resolution videos, and manually annotated with a bounding box and a label describing 10 different visibility levels, categorized according to the percentage of the human body visible inside the bounding box. This allows computer vision models to be evaluated on their detection performance across different ranges of occlusion. NOMAD is designed to improve the effectiveness of aerial search and rescue and to enhance collaboration between sUAS and humans, by providing a new benchmark dataset for human detection under occluded aerial views.

****

UNSPAT: Uncertainty-Guided SpatioTemporal Transformer for 3D Human Pose and Shape Estimation on Videos

Minsoo Lee, Hyunmin Lee, Bumsoo Kim, Seunghwan Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3004-3013

We propose an efficient framework for 3D human pose and shape estimation from a video, named Uncertainty-Guided SpatioTemporal Transformer (UNSPAT). Unlike previous video-based methods that consider temporal relationships with global average pooled features, our approach incorporates both spatial and temporal dimensions without compromising spatial information. We address the excessive complexity of spatiotemporal attention through two modules: Spatial Alignment Module (SAM) and Space2Batch. The modules align input features and compute temporal attention at every spatial position in a batch-wise manner. Furthermore, our uncertainty-guided attention re-weighting module improves performance by diminishing the impact of artifacts. We demonstrate the effectiveness of the UNSPAT on widely used benchmark datasets and achieve state-of-the-art performance. Our method is robus to challenging scenes, such as occlusion, and cluttered backgrounds, showing its potential for real-world applications.

****

Consistent Multimodal Generation via a Unified GAN Framework
Zhen Zhu, Yijun Li, Weijie Lyu, Krishna Kumar Singh, Zhixin Shu, Sören Pirk, Derek Hoiem; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5048-5057
We investigate how to generate multimodal image outputs, such as RGB, depth, and surface normals, with a single generative model. The challenge is to produce outputs that are realistic, and also consistent with each other. Our solution builds on the StyleGAN3 architecture, with a shared backbone and modality-specific branches in the last layers of the synthesis network, and we propose per-modality fidelity discriminators and a cross-modality consistency discriminator. In experiments on the Stanford2D3D dataset, we demonstrate realistic and consistent generation of RGB, depth, and normal images. We also show a training recipe to easily extend our pretrained model on a new domain, even with a few pairwise data. We further evaluate the use of synthetically generated RGB and depth pairs for training or fine-tuning depth estimators. Code will be available at https://github.com/jessemelpolio/MultimodalGAN.

****

Self-Supervised Learning of Semantic Correspondence Using Web Videos
Donghyeon Kwon, Minsu Cho, Suha Kwak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2142-2152
Existing datasets for semantic correspondence are often limited in terms of both the amount of labeled data and diversity of labeled keypoints due to the tremendous cost of manual correspondence labeling. To address this issue, we propose the first self-supervised learning framework that utilizes a large amount of web videos collected and annotated fully automatically. Our main motivation is that smooth changes between consecutive video frames allow to build accurate space-time correspondences with no human intervention. Hence, we establish space-time correspondences within each web video and leverage them for deriving pseudo correspondence labels between two distant frames of the video. In addition, we present a dedicated training strategy that facilitates stable training using web videos with such pseudo labels. Our experiments on public benchmarks demonstrated that the proposed method surpasses existing self-supervised learning models and that our self-supervised learning as pretraining for supervised learning improves performance substantially. Our codebase for web video crawling and pseudo label generation will be released public to promote future research.

****

TIAM - A Metric for Evaluating Alignment in Text-to-Image Generation
Paul Grimal, Hervé Le Borgne, Olivier Ferret, Julien Tourille; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2890-2899
The progress in the generation of synthetic images has made it crucial to assess their quality. While several metrics have been proposed to assess the rendering of images, it is crucial for Text-to-Image (T2I) models, which generate images based on a prompt, to consider additional aspects such as to which extent the generated image matches the important content of the prompt. Moreover, although the generated images usually result from a random starting point, the influence of this one is generally not considered. In this article, we propose a new metric based on prompt templates to study the alignment between the content specified in the prompt and the corresponding generated images. It allows us to better characterize the alignment in terms of the type of the specified objects, their number, and their color. We conducted a study on several recent T2I models about various aspects. An additional interesting result we obtained with our approach is that image quality can vary drastically depending on the noise used as a seed for the images. We also quantify the influence of the number of concepts in the prompt, their order as well as their (color) attributes. Finally, our method allows us to identify some seeds that produce better images than others, opening novel directions of research on this understudied topic.

****

HDMNet: A Hierarchical Matching Network With Double Attention for Large-Scale Outdoor LiDAR Point Cloud Registration

Weiyi Xue, Fan Lu, Guang Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3393-3403

Outdoor LiDAR point clouds are typically large-scale and complexly distributed. To achieve efficient and accurate registration, emphasizing the similarity among local regions and prioritizing global local-to-local matching is of utmost importance, subsequent to which accuracy can be enhanced through cost-effective fine registration. In this paper, a novel hierarchical neural network with double attention named HDMNet is proposed for large-scale outdoor LiDAR point cloud registration. Specifically, A novel feature consistency enhanced double-soft matching network is introduced to achieve two-stage matching with high flexibility while enlarging the receptive field with high efficiency in a patch-to-patch manner, which significantly improves the registration performance. Moreover, in order to further utilize the sparse matching information from deeper layer, we develop a novel trainable embedding mask to incorporate the confidence scores of correspondences obtained from pose estimation of deeper layer, eliminating additional computations. The high-confidence keypoints in the sparser point cloud of the deeper layer correspond to a high-confidence spatial neighborhood region in shallower layer, which will receive more attention, while the features of non-key regions will be masked. Extensive experiments are conducted on two large-scale outdoor LiDAR point cloud datasets to demonstrate the high accuracy and efficiency of the proposed HDMNet.

****

UGPNet: Universal Generative Prior for Image Restoration

Hwayoon Lee, Kyoungkook Kang, Hyeongmin Lee, Seung-Hwan Baek, Sunghyun Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1598-1608

Recent image restoration methods can be broadly categorized into two classes: (1) regression methods that recover the rough structure of the original image without synthesizing high-frequency details and (2) generative methods that synthesize perceptually-realistic high-frequency details even though the resulting image deviates from the original structure of the input. While both directions have been extensively studied in isolation, merging their benefits with a single framework has been rarely studied. In this paper, we propose UGPNet, a universal image restoration framework that can effectively achieve the benefits of both approaches by simply adopting a pair of an existing regression model and a generative model. UGPNet first restores the image structure of a degraded input using a regression model and synthesizes a perceptually-realistic image with a generative model on top of the regressed output. UGPNet then combines the regressed output and the synthesized output, resulting in a final result that faithfully reconstructs the structure of the original image in addition to perceptually-realistic textures. Our extensive experiments on deblurring, denoising, and super-resolution demonstrate that UGPNet can successfully exploit both regression and generative methods for high-fidelity image restoration.

****

Defense Against Adversarial Cloud Attack on Remote Sensing Salient Object Detection

Huiming Sun, Lan Fu, Jinlong Li, Qing Guo, Zibo Meng, Tianyun Zhang, Yuewei Lin, Hongkai Yu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8345-8354

Detecting the salient objects in a remote sensing image has wide applications. Many existing deep learning methods have been proposed for Salient Object Detection (SOD) in remote sensing images with remarkable results. However, the recent adversarial attack examples, generated by changing a few pixel values on the original image, could result in a collapse for the well-trained deep learning model. Different with existing methods adding perturbation to original images, we propose to jointly tune adversarial exposure and additive perturbation for attack and constrain image close to cloudy image as Adversarial Cloud. Cloud is natural a nd common in remote sensing images, however, camouflaging cloud based adversarial attack and defense for remote sensing images are not well studied before. Furthermore, we design DefenseNet as a learnable pre-processing to the adversarial cloudy images to preserve the performance of the deep learning based remote sensing SOD model, without tuning the already deployed deep SOD model. By considering both regular and generalized adversarial examples, the proposed DefenseNet can defend the proposed Adversarial Cloud in white-box setting and other attack methods in black-box setting. Experimental results on a synthesized benchmark from the public remote sensing dataset (EORSSD) show the promising defense against adversarial cloud attacks.

****

Diffusion in the Dark: A Diffusion Model for Low-Light Text Recognition

Cindy M. Nguyen, Eric R. Chan, Alexander W. Bergman, Gordon Wetzstein; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4146-4157

Capturing images is a key part of automation for high-level tasks such as scene text recognition. Low-light conditions pose a challenge for high-level perception stacks, which are often optimized on well-lit, artifact-free images. Reconstruction methods for low-light images can produce well-lit counterparts, but typically at the cost of high-frequency details critical for downstream tasks. We propose Diffusion in the Dark (DiD), a diffusion model for low-light image reconstruction for text recognition. DiD provides qualitatively competitive reconstructions with that of state-of-the-art (SOTA), while preserving high-frequency details even in extremely noisy, dark conditions. We demonstrate that DiD, without any task-specific optimization, can outperform SOTA low-light methods in low-light text recognition on real images, bolstering the potential of diffusion models to solve ill-posed inverse problems.

****

RobustCLEVR: A Benchmark and Framework for Evaluating Robustness in Object-Centric Learning

Nathan Drenkow, Mathias Unberath; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4518-4527

Object-centric representation learning offers the potential to overcome limitations of image-level representations by explicitly parsing image scenes into their constituent components. While image-level representations typically lack robustness to natural image corruptions, the robustness of object-centric methods remains largely untested. To address this gap, we present the RobustCLEVR benchmark dataset and evaluation framework. Our framework takes a novel approach to evaluating robustness by enabling the specification of causal dependencies in the image generation process grounded in expert knowledge and capable of producing a wide range of image corruptions unattainable in existing robustness evaluations. Using our framework, we define several causal models of the image corruption process which explicitly encode assumptions about the causal relationships and distributions of each corruption type. We generate dataset variants for each causal model on which we evaluate state-of-the-art object-centric methods. Overall, we find that object-centric methods are not inherently robust to image corruptions. Our causal evaluation approach exposes model sensitivities not observed using conventional evaluation processes, yielding greater insight into robustness differences across algorithms. Lastly, while conventional robustness evaluations view corruptions as out-of-distribution, we use our causal framework to show that even training on in-distribution image corruptions does not guarantee increased model robustness. This work provides a step towards more concrete and substantiated understanding of model performance and deterioration under complex corruption processes of the real-world.

****

AFTer-SAM: Adapting SAM With Axial Fusion Transformer for Medical Imaging Segmentation

Xiangyi Yan, Shanlin Sun, Kun Han, Thanh-Tung Le, Haoyu Ma, Chenyu You, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7975-7984

The Segmentation Anything Model (SAM) has demonstrated effectiveness in various segmentation tasks. However, its application to 3D medical data has posed challenges due to its inherent design for both 2D and natural images. While there have been attempts to apply SAM to medical images on a slice-by-slice basis, the outcomes have been less than optimal. In this study, we introduce AFTer-SAM, an adaptation of SAM designed for volumetric medical image segmentation. By incorporating an Axial Fusion Transformer, AFTer-SAM is capable of capturing both intra-slice details and inter-slice contextual information, essential for accurate medical image segmentation. Given the potential computational challenges of training this enhanced model, we utilize Low-Rank Adaptation (LoRA) to efficiently fine-tune the weights of the Axial Fusion Transformer. This ensures a streamlined training process without compromising on performance. Our results indicate that AFTer-SAM offers significant improvements in volumetric medical image segmentation, suggesting a promising direction for the application of large pre-trained models in medical imaging.

****

Plasticity-Optimized Complementary Networks for Unsupervised Continual Learning

Alex Gomez-Villa, Bartlomiej Twardowski, Kai Wang, Joost van de Weijer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1690-1700

Continuous unsupervised representation learning (CURL) research has greatly benefited from improvements in self-supervised learning (SSL) techniques. As a result, existing CURL methods using SSL can learn high-quality representations without any labels, but with a notable performance drop when learning on a many-tasks data stream. We hypothesize that this is caused by the regularization losses that are imposed to prevent forgetting, leading to a suboptimal plasticity-stability trade-off: they either do not adapt fully to the incoming data (low plasticity), or incur significant forgetting when allowed to fully adapt to a new SSL pretext-task (low stability). In this work, we propose to train an expert network that is relieved of the duty of keeping the previous knowledge and can focus on performing optimally on the new tasks (optimizing plasticity). In the second phase, we combine this new knowledge with the previous network in an adaptation-retrospection phase to avoid forgetting and initialize a new expert with the knowledge of the old network. We perform several experiments showing that our proposed approach outperforms other CURL exemplar-free methods in few- and many-task split settings. Furthermore, we show how to adapt our approach to semi-supervised continual learning (Semi-SCL) and show that we surpass the accuracy of other exemplar-free Semi-SCL methods and reach the results of some others that use exemplars.

****

FATE: Feature-Agnostic Transformer-Based Encoder for Learning Generalized Embedding Spaces in Flow Cytometry Data

Lisa Weijler, Florian Kowarsch, Michael Reiter, Pedro Hermosilla, Margarita Maurer-Granofszky, Michael Dworzak; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7956-7964

While model architectures and training strategies have become more generic and flexible with respect to different data modalities over the past years, a persistent limitation lies in the assumption of fixed quantities and arrangements of input features. This limitation becomes particularly relevant in scenarios where the attributes captured during data acquisition vary across different samples. In this work, we aim at effectively leveraging data with varying features, without the need to constrain the input space to the intersection of potential feature sets or to expand it to their union. We propose a novel architecture that can directly process data without the necessity of aligned feature modalities by learning a general embedding space that captures the relationship between features across data samples with varying sets of features. This is achieved via a set-transformer architecture augmented by feature-encoder layers, thereby enabling the learning of a shared latent feature space from data originating from heterogeneous feature spaces. The advantages of the model are demonstrated for automatic cancer cell detection in acute myeloid leukemia in flow cytometry data, where the f eatures measured during acquisition often vary between samples. Our proposed architecture's capacity to operate seamlessly across incongruent feature spaces is particularly relevant in this context, where data scarcity arises from the low prevalence of the disease. The code is available for research purposes at https://github.com/lisaweijler/FATE.

****

Label Augmentation As Inter-Class Data Augmentation for Conditional Image Synthesis With Imbalanced Data

Kai Katsumata, Duc Minh Vo, Hideki Nakayama; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4944-4953

Conditional image synthesis performs admirably when trained on well-constructed and balanced datasets. However, in practice, training datasets frequently contain minorities (i.e., a class with a few samples), known as imbalanced data, which causes difficulties in learning generative models. To address conditional image synthesis with imbalanced data, we analyze a diversity issue of label-preserving data augmentation and an affinity issue of non-label-preserving data augmentation. From this observation, we present label augmentation, which works as inter-class data augmentation that effectively augments data by predicting a new label for a given image using the prediction of a pretrained image classification model (i.e., probabilities for each class). We incorporate our label augmentation into the discriminator of a seminal conditional generative adversarial network (GAN) model, proposing Softlabel-GAN. Using class probabilities extracts class-invariant and shared features between similar classes, achieving data augmentation with high affinity and diversity. Our experiments on imbalanced datasets show that Softlabel-GAN produces images with high quality and diversity while being hardly affected by the number of samples in each class. Code: https://github.com/raven38/softlabel-gan.

****

Sign Language Production With Latent Motion Transformer

Pan Xie, Taiying Peng, Yao Du, Qipeng Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3024-3034

Sign Language Production (SLP) is the tough task of turning sign language into sign videos. The main goal of SLP is to create these videos using a sign gloss. In this research, we've developed a new method to make high-quality sign videos without using human poses as a middle step. Our model works in two main parts: first, it learns from a generator and the video's hidden features, and next, it uses another model to understand the order of these hidden features. To make this method even better for sign videos, we make several significant improvements. (i) In the first stage, we take an improved 3D VQ-GAN to learn downsampled latent representations. (ii) In the second stage, we introduce sequence-to-sequence attention to better leverage conditional information. (iii) The separated two-stage training discards the realistic visual semantic of the latent codes in the second stage. To endow the latent sequences semantic information, we extend the token-level autoregressive latent codes learning with perceptual loss and reconstruction loss for the prior model with visual perception. Compared with previous state-of-the-art approaches, our model performs consistently better on two word-level sign language datasets, i.e., WLASL and NMFs-CSL.

****

Diffused Heads: Diffusion Models Beat GANs on Talking-Face Generation

Micha■ Stypu■kowski, Konstantinos Vougioukas, Sen He, Maciej Zi■ba, Stavros Petridis, Maja Pantic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5091-5100

Talking face generation has historically struggled to produce head movements and natural facial expressions without guidance from additional reference videos. Recent developments in diffusion-based generative models allow for more realistic and stable data synthesis and their performance on image and video generation has surpassed that of other generative models. In this work, we present an autoregressive diffusion model that requires only one identity image and audio sequence to generate a video of a realistic talking head. Our solution is capable of hallucinating head movements, facial expressions, such as blinks, and preserving a

given background. We evaluate our model on two different datasets, achieving state-of-the-art results in expressiveness and smoothness on both of them.
****

U3DS3: Unsupervised 3D Semantic Scene Segmentation
Jiaxu Liu, Zhengdi Yu, Toby P. Breckon, Hubert P. H. Shum; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3759-3768

Contemporary point cloud segmentation approaches largely rely on richly annotated 3D training data. However, it is both time-consuming and challenging to obtain consistently accurate annotations for such 3D scene data. Moreover, there is still a lack of investigation into fully unsupervised scene segmentation for point clouds, especially for holistic 3D scenes. This paper presents U3DS3, as a step towards completely unsupervised point cloud segmentation for any holistic 3D scenes. To achieve this, U3DS3 leverages a generalized unsupervised segmentation method for both object and background across both indoor and outdoor static 3D point clouds with no requirement for model pre-training, by leveraging only the inherent information of the point cloud to achieve full 3D scene segmentation. The initial step of our proposed approach involves generating superpoints based on the geometric characteristics of each scene. Subsequently, it undergoes a learning process through a spatial clustering-based methodology, followed by iterative training using pseudo-labels generated in accordance with the cluster centroids. Moreover, by leveraging the invariance and equivariance of the volumetric representations, we apply the geometric transformation on voxelized features to provide two sets of descriptors for robust representation learning. Finally, our evaluation provides state-of-the-art results on the ScanNet and SemanticKITTI, and competitive results on the S3DIS, benchmark datasets.
****

GIPCOL: Graph-Injected Soft Prompting for Compositional Zero-Shot Learning
Guangyue Xu, Joyce Chai, Parisa Kordjamshidi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5774-5783

Pre-trained vision-language models (VLMs) have achieved promising success in many fields, specially with prompt learning paradigm. However, designing proper textual prompts to adapt VLMs for downstream tasks is still challenging. In this work, we propose GIPCOL (Graph-Injected soft Prompting for COmpositional Learning) to better explore the compositional zero-shot learning (CZSL) ability of VLMs within the prompt-based learning framework. The soft prompt in GIPCOL is structured and consists of the prefix learnable vectors, attribute label and object label. In addition, the attribute and object labels in the soft prompt are designated as nodes in a compositional graph. The compositional graph is constructed based on the compositional structure of the objects and attributes extracted from the training data and consequently feeds the updated concept representation into the soft prompt to capture this compositional structure for a better CZSL learning. With the new prompting strategy, GIPCOL achieves state-of-the-art AUC results on all three CZSL benchmarks, including MIT-States, UT-Zappos, and C-GQA datasets in both closed and open settings compared to previous non-CLIP as well as CLIP-based methods. We analyze when and why GIPCOL operates well given the CLIP backbone and its training data limitations, and our findings shed light on designing prompts for CZSL.
****

STEP - Towards Structured Scene-Text Spotting
Sergi Garcia-Bordils, Dimosthenis Karatzas, Marçal Rusiñol; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 883-892

We introduce the structured scene-text spotting task, which requires a scene-text OCR system to spot text in the wild according to a query regular expression. Contrary to generic scene-text OCR, structured scene-text spotting seeks to dynamically condition both detection and recognition on user-provided regular expressions. To tackle this task, we propose the Structured TExt sPotter (STEP), a model that exploits the provided text structure to guide the OCR process. STEP is able to deal with regular expressions that contain spaces and it is not bound to d etection at word-level granularity. Our approach enables accurate zero-shot structured text spotting in a wide variety of real-world reading scenarios and is solely trained on publicly available data. To demonstrate the effectiveness of our approach, we introduce a new challenging test dataset that contains several types of out-of-vocabulary structured text, reflecting important reading applications such as weight information, serial numbers, license plates etc. We demonstrate that STEP can provide specialized OCR performance on demand in all tested scenarios. The code and test dataset are released at https://github.com/CVC-DAG/STEP.

****

ClipSitu: Effectively Leveraging CLIP for Conditional Predictions in Situation Recognition

Debaditya Roy, Dhruv Verma, Basura Fernando; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 444-453

Situation Recognition is the task of generating a structured summary of what is happening in an image using an activity verb and the semantic roles played by actors and objects. In this task, the same activity verb can describe a diverse set of situations as well as the same actor or object category can play a diverse set of semantic roles depending on the situation depicted in the image. Hence a situation recognition model needs to understand the context of the image and the visual-linguistic meaning of semantic roles. Therefore, we leverage the CLIP foundational model that has learned the context of images via language descriptions. We show that deeper-and-wider multi-layer perceptron (MLP) blocks obtain noteworthy results for the situation recognition task by using CLIP image and text embedding features and it even outperforms the state-of-the-art CoFormer, a Transformer-based model, thanks to the external implicit visual-linguistic knowledge encapsulated by CLIP and the expressive power of modern MLP block designs. Motivated by this, we design a cross-attention-based Transformer using CLIP visual tokens that model the relation between textual roles and visual entities. Our cross-attention-based Transformer known as ClipSitu XTF outperforms existing state-of-the-art by a large margin of 14.1% on semantic role labelling (value) for top-1 accuracy using imSitu dataset. Similarly, our ClipSitu XTF obtains state-of-the-art situation localization performance. We will make the code publicly available.

****

Multimodality-Guided Image Style Transfer Using Cross-Modal GAN Inversion

Hanyu Wang, Pengxiang Wu, Kevin Dela Rosa, Chen Wang, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4976-4985

Image Style Transfer (IST) is an interdisciplinary topic of computer vision and art that continuously attracts researchers' interests. Different from traditional Image-guided Image Style Transfer (IIST) methods that require a style reference image as input to define the desired style, recent works start to tackle the problem in a text-guided manner, i.e., Text-guided Image Style Transfer (TIST). Compared to IIST, such approaches provide more flexibility with text-specified styles, which are useful in scenarios where the style is hard to define with reference images. Unfortunately, many TIST approaches produce undesirable artifacts in the transferred images. To address this issue, we present a novel method to achieve much improved style transfer based on text guidance. Meanwhile, to offer more flexibility than IIST and TIST, our method allows style inputs from multiple sources and modalities, enabling MultiModality-guided Image Style Transfer (MMIST). Specifically, we realize MMIST with a novel cross-modal GAN inversion method, which generates style representations consistent with specified styles. Such style representations facilitate style transfer and in principle generalize any IIST methods to MMIST. Large-scale experiments and user studies demonstrate that our method achieves state-of-the-art performance on TIST task. Furthermore, comprehensive qualitative results confirm the effectiveness of our method on MMIST task and cross-modal style interpolation.

****

Meta-Learned Attribute Self-Interaction Network for Continual and Generalized Ze

ro-Shot Learning
Vinay Verma, Nikhil Mehta, Kevin J. Liang, Aakansha Mishra, Lawrence Carin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2721-2731

Zero-shot learning (ZSL) is a promising approach to generalizing a model to categories unseen during training by leveraging class attributes, but challenges remain. Recently, methods using generative models to combat bias towards classes seen during training have pushed state of the art, but these generative models can be slow or computationally expensive to train. Also, these generative models assume that the attribute vector of each unseen class is available a priori at training, which is not always practical. Additionally, while many previous ZSL methods assume a one-time adaptation to unseen classes, in reality, the world is always changing, necessitating a constant adjustment of deployed models. Models unprepared to handle a sequential stream of data are likely to experience catastrophic forgetting. We propose a Meta-learned Attribute self-Interaction Network (MAIN) for continual ZSL. By pairing attribute self-interaction trained using meta-learning with inverse regularization of the attribute encoder, we are able to outperform state-of-the-art results without leveraging the unseen class attributes while also being able to train our models substantially faster (>100x) than expensive generative-based approaches. We demonstrate this with experiments on five standard ZSL datasets (CUB, aPY, AWA1, AWA2, and SUN) in the generalized zero-shot learning and continual (fixed/dynamic) zero-shot learning settings. Extensive ablations and analyses demonstrate the efficacy of various components proposed.

****

MoP-CLIP: A Mixture of Prompt-Tuned CLIP Models for Domain Incremental Learning
Julien Nicolas, Florent Chiaroni, Imtiaz Ziko, Ola Ahmad, Christian Desrosiers, Jose Dolz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1762-1772

Despite the recent progress in incremental learning, addressing catastrophic forgetting under distributional drift is still an open and important problem. Indeed, while state-of-the-art domain incremental learning (DIL) methods perform satisfactorily within known domains, their performance largely degrades in the presence of novel domains. This limitation hampers their generalizability, and restricts their scalability to more realistic settings where train and test data are drawn from different distributions. To address these limitations, we present a novel DIL approach based on a mixture of prompt-tuned CLIP models (MoP-CLIP), which generalizes the paradigm of S-Prompting to handle both in-distribution and out-of-distribution data at inference. In particular, at the training stage we model the features distribution of every class in each domain, learning individual text and visual prompts to adapt to a given domain. At inference, the learned distributions allow us to identify whether a given test sample belongs to a known domain, selecting the correct prompt for the classification task, or from an unseen domain, leveraging a mixture of the prompt-tuned CLIP models. Our empirical evaluation reveals the poor performance of existing DIL methods under domain shift, and suggests that the proposed MoP-CLIP performs competitively in the standard DIL settings while outperforming state-of-the-art methods in OOD scenarios. These results demonstrate the superiority of MoP-CLIP, offering a robust and general solution to the problem of domain incremental learning.

****

So You Think You Can Track?
Derek Gloudemans, Gergely Zachár, Yanbing Wang, Junyi Ji, Matt Nice, Matt Bunting, William W. Barbour, Jonathan Sprinkle, Benedetto Piccoli, Maria Laura Delle Monache, Alexandre Bayen, Benjamin Seibold, Daniel B. Work; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4528-4538

This work introduces a multi-camera tracking dataset consisting of 234 hours of video data recorded concurrently from 234 overlapping HD cameras covering a 4.2 mile stretch of 8-10 lane interstate highway near Nashville, TN. The video is recorded during a period of high traffic density with 500+ objects typically visib le within the scene and typical object longevities of 3-15 minutes. GPS trajectories from 270 vehicle passes through the scene are manually corrected in the video data to provide a set of ground-truth trajectories for recall-oriented tracking metrics, and object detections are provided for each camera in the scene (159 million total before cross-camera fusion). Initial benchmarking of tracking-by-detection algorithms is performed against the GPS trajectories, and a best HOTA of only 9.5% is obtained (best recall 75.9% at IOU 0.1, 47.9 average IDs per ground truth object), indicating the benchmarked trackers do not perform sufficiently well at the long temporal and spatial durations required for traffic scene understanding.

****

OmniVec: Learning Robust Representations With Cross Modal Sharing

Siddharth Srivastava, Gaurav Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1236-1248

Majority of research in learning based methods has been towards designing and training networks for specific tasks. However, many of the learning based tasks, across modalities, share commonalities and could be potentially tackled in a joint framework. We present an approach in such direction, to learn multiple tasks, in multiple modalities, with a unified architecture. The proposed network is composed of task specific encoders, a common trunk in the middle, followed by task specific prediction heads. We first pre-train it by self-supervised masked training, followed by sequential training for the different tasks. We train the network on all major modalities, e.g. visual, audio, text and 3D, and report results on 22 diverse and challenging public benchmarks. We demonstrate empirically that, using a joint network to train across modalities leads to meaningful information sharing and this allows us to achieve state-of-the-art results on most of the benchmarks. We also show generalization of the trained network on cross-modal tasks as well as unseen datasets and tasks.

****

MSCC: Multi-Scale Transformers for Camera Calibration

Xu Song, Hao Kang, Atsunori Moteki, Genta Suzuki, Yoshie Kobayashi, Zhiming Tan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3262-3271

Camera calibration is very important for some vision tasks, like rendering 3D scenes, environment reconstruction, and self-localization, etc. In this paper, we propose a framework of multi-scale transformers for camera calibration. With the input of a single image, the multi-scale features output from the model's backbone are utilized to estimate camera parameters. At the same time, we show that the way of coarse-to-fine is effective to locate global structures and detailed features in the image, by studying the attention response of horizon line estimation. Moreover, deep supervision is proven to get more precise results and accelerated training. Our method outperforms all the state-of-the-art methods by objective and subjective experiments on Google Street View dataset and Pano360.

****

Multi-Modal Gaze Following in Conversational Scenarios

Yuqi Hou, Zhongqun Zhang, Nora Horanyi, Jaewon Moon, Yihua Cheng, Hyung Jin Chang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1186-1195

Gaze following estimates gaze targets of in-scene person by understanding human behavior and scene information. Existing methods usually analyze scene images for gaze following. However, compared with visual images, audio also provides crucial cues for determining human behavior.This suggests that we can further improve gaze following considering audio cues. In this paper, we explore gaze following tasks in conversational scenarios. We propose a novel multi-modal gaze following framework based on our observation "audiences tend to focus on the speaker". We first leverage the correlation between audio and lips, and classify speakers and listeners in a scene. We then use the identity information to enhance scene images and propose a gaze candidate estimation network. The network estimates gaze candidates from enhanced scene images and we use MLP to match subjects with candidates as classification tasks. Existing gaze following datasets focus on vis ual images while ignore audios.To evaluate our method, we collect a conversational dataset, VideoGazeSpeech (VGS), which is the first gaze following dataset including images and audio. Our method significantly outperforms existing methods in VGS datasets. The visualization result also prove the advantage of audio cues in gaze following tasks. Our work will inspire more researches in multi-modal gaze following estimation.

****

Contrastive Viewpoint-Aware Shape Learning for Long-Term Person Re-Identification

Vuong D. Nguyen, Khadija Khaldi, Dung Nguyen, Pranav Mantini, Shishir Shah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1041-1049

Traditional approaches for Person Re-identification (Re-ID) rely heavily on modeling the appearance of persons. This measure is unreliable over longer durations due to the possibility for changes in clothing or biometric information. Furthermore, viewpoint changes significantly degrade the matching ability of these methods. In this paper, we propose "Contrastive Viewpoint-aware Shape Learning for Long-term Person Re-Identification" (CVSL) to address these challenges. Our method robustly extracts local and global texture-invariant human body shape cues from 2D pose using the Relational Shape Embedding branch, which consists of a pose estimator and a shape encoder built on a Graph Attention Network. To enhance the discriminability of the shape and appearance of identities under viewpoint variations, we propose Contrastive Viewpoint-aware Losses (CVL). CVL leverages contrastive learning to simultaneously minimize the intra-class gap under different viewpoints and maximize the inter-class gap under the same viewpoint. Extensive experiments demonstrate that our proposed framework outperforms state-of-the-art methods on long-term person Re-ID benchmarks.

****

Scale-Adaptive Feature Aggregation for Efficient Space-Time Video Super-Resolution

Zhewei Huang, Ailin Huang, Xiaotao Hu, Chen Hu, Jun Xu, Shuchang Zhou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4228-4239

The Space-Time Video Super-Resolution (STVSR) task aims to enhance the visual quality of videos, by simultaneously performing video frame interpolation (VFI) and video super-resolution (VSR). However, facing the challenge of the additional temporal dimension and scale inconsistency, most existing STVSR methods are complex and inflexible in dynamically modeling different motion amplitudes. In this work, we find that choosing an appropriate processing scale achieves remarkable benefits in flow-based feature propagation. We propose a novel Scale-Adaptive Feature Aggregation (SAFA) network that adaptively selects sub-networks with different processing scales for individual samples. Experiments on four public STVSR benchmarks demonstrate that SAFA achieves state-of-the-art performance. Our SAFA network outperforms recent state-of-the-art methods such as TMNet and VideoINR by an average improvement of over 0.5dB on PSNR, while requiring less than half the number of parameters and only 1/3 computational costs. Our code will be publicly released.

****

SSP: Semi-Signed Prioritized Neural Fitting for Surface Reconstruction From Unoriented Point Clouds

Runsong Zhu, Di Kang, Ka-Hei Hui, Yue Qian, Shi Qiu, Zhen Dong, Linchao Bao, Pheng-Ann Heng, Chi-Wing Fu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3769-3778

Reconstructing 3D geometry from unoriented point clouds can benefit many downstream tasks. % Recent shape modeling methods mostly adopt implicit neural representation to fit a signed distance field (SDF) and optimize the network by unsigned supervision. % However, these methods occasionally have difficulty in finding the coarse shape for complicated objects, especially suffering from the "ghost" surfaces (i.e., fake surfaces that should not exist). % To guide the network quickly fit the coarse shape, we propose to utilize the signed supervision in region s that are obviously outside the object and can be easily determined, resulting in our semi-signed supervision. % To better recover high-fidelity details, a novel loss-based region sampling strategy and a progressive positional encoding (PE) method are applied to prioritize the optimization towards underfitting and complicated regions. % Specifically, we voxelize and partition the object space into sign-known and sign-uncertain regions, in which different supervisions are applied. % Besides, we adaptively adjust the sampling rate of each voxel according to the tracked reconstruction loss, so that the network can focus more on the complicated under-fitting regions. % We conduct extensive experiments to demonstrate that our method achieves state-of-the-art performance compared to the existing fitting-based methods and comparable performance to learning-based methods on multiple datasets. % The code is publicly available at https://github.com/Runsong123/SSP.

****

DeVos: Flow-Guided Deformable Transformer for Video Object Segmentation

Volodymyr Fedynyak, Yaroslav Romanus, Bohdan Hlovatskyi, Bohdan Sydor, Oles Dobosevych, Igor Babin, Roman Riazantsev; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 240-249

The recent works on Video Object Segmentation achieved remarkable results by matching dense semantic and instance-level features between the current and previous frames for long-time propagation. Nevertheless, global feature matching ignores scene motion context, failing to satisfy temporal consistency. Even though some methods introduce local matching branch to achieve smooth propagation, they fail to model complex appearance changes due to the constraints of the local window. In this paper, we present DeVOS (Deformable VOS), an architecture for Video Object Segmentation that combines memory-based matching with motion-guided propagation resulting in stable long-term modeling and strong temporal consistency. For short-term local propagation, we propose a novel attention mechanism ADVA (Adaptive Deformable Video Attention), allowing the adaption of similarity search region to query-specific semantic features, which ensures robust tracking of complex shape and scale changes. DeVOS employs an optical flow to obtain scene motion features which are further injected to deformable attention as strong priors to learnable offsets. Our method achieves top-rank performance on DAVIS 2017 val and test-dev (88.1%, 83.0%), YouTube-VOS 2019 val (86.6%) while featuring consistent run-time speed and stable memory consumption.

****

GraphFill: Deep Image Inpainting Using Graphs

Shashikant Verma, Aman Sharma, Roopa Sheshadri, Shanmuganathan Raman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4996-5006

We present a novel coarser-to-finer approach for deep graphical image inpainting that utilizes GraphFill, a graph neural network-based deep learning framework, and a lightweight generative baseline network. We construct a pyramidal graph for the input-masked image by reducing it into superpixels, each representing a node in the graph. The proposed pyramidal approach facilitates the transfer of global context from coarser to finer pyramid levels, enabling GraphFill to estimate plausible information for unknown node values in the graph. The estimated information is used to fill in the masked region, which a Refine Network then refines. Furthermore, we propose a resolution-robust pyramidal graph construction method, allowing for efficient inpainting of high-resolution images with relatively fewer computations. Our proposed network, trained on Places and CelebA-HQ datasets, demonstrates competitive performance compared to existing methods while using fewer learning parameters. We conduct thorough ablation studies to evaluate the effectiveness of each component in the GraphFill Network for improved performance. Our proposed lightweight model for image inpainting is efficient in real-world scenarios, as it can be easily deployed on mobile devices with limited resources.

****

AU-Aware Dynamic 3D Face Reconstruction From Videos With Transformer

Chenyi Kuang, Jeffrey O. Kephart, Qiang Ji; Proceedings of the IEEE/CVF Winter C onference on Applications of Computer Vision (WACV), 2024, pp. 6237-6247

In spite of the significant progresses in monocular or multi-view image based 3D face reconstruction research, recovering 3D faces from videos, which contains rich dynamic information of facial motions, still remains as a highly challenging problem. First, most prior works fail to generate accurate and stable 3D faces on videos, especially for recovering subtle expression details. Furthermore, existing dynamic reconstruction approaches have not fully considered the temporal dependency of facial expression transitions, which is based on the dynamic muscle activation system under a local region of the skin. To tackle the aforementioned challenges, we present a framework for dynamic 3D face reconstruction from monocular videos, which can accurately recover 3D facial geometrical representations for facial action unit (AU). Specifically, we design a coarse-to-fine framework, where the "coarse" 3D face sequences are generated by a pre-trained static reconstruction model; and the "refinement" is performed through a Transformer-based network. We design 1) a Temporal Module used for modeling temporal dependency of facial motion dynamics; 2) an Spatial Module for modeling AU spatial correlations from geometry-based AU tokens; 3) feature fusion for simultaneous dynamic facial AU recognition and 3D expression capturing. Experimental results show the superiority of our method in generating AU-aware 3D face reconstruction sequences both quantitatively and qualitatively.

****

Unified Concept Editing in Diffusion Models

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzy■ska, David Bau; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5111-5120

Text-to-image models suffer from various safety issues that may limit their suitability for deployment. Previous methods have separately addressed individual issues of bias, copyright, and offensive content in text-to-image models. However, in the real world, all of these issues appear simultaneously in the same model. We present a method that tackles all issues with a single approach. Our method, Unified Concept Editing (UCE), edits the model without training using a closed-form solution, and scales seamlessly to concurrent edits on text-conditional diffusion models. We demonstrate scalable simultaneous debiasing, style erasure, and content moderation by editing text-to-image projections, and we present extensive experiments demonstrating improved efficacy and scalability over prior work.

****

MEGANet: Multi-Scale Edge-Guided Attention Network for Weak Boundary Polyp Segmentation

Nhat-Tan Bui, Dinh-Hieu Hoang, Quang-Thuc Nguyen, Minh-Triet Tran, Ngan Le; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7985-7994

Efficient polyp segmentation in healthcare plays a critical role in enabling early diagnosis of colorectal cancer. However, the segmentation of polyps presents numerous challenges, including the intricate distribution of backgrounds, variations in polyp sizes and shapes, and indistinct boundaries. Defining the boundary between the foreground (i.e. polyp itself) and the background (surrounding tissue) is difficult. To mitigate these challenges, we propose Multi-Scale Edge-Guided Attention Network (MEGANet) tailored specifically for polyp segmentation within colonoscopy images. This network draws inspiration from the fusion of a classical edge detection technique with an attention mechanism. By combining these techniques, MEGANet effectively preserves high-frequency information, notably edges and boundaries, which tend to erode as neural networks deepen. MEGANet is designed as an end-to-end framework, encompassing three key modules: an encoder, which is responsible for capturing and abstracting the features from the input image, a decoder, which focuses on salient features, and the Edge-Guided Attention module (EGA) that employs the Laplacian Operator to accentuate polyp boundaries. Extensive experiments, both qualitative and quantitative, on five benchmark datasets, demonstrate that our MEGANet outperforms other existing SOTA methods under six evaluation metrics. Our code is available at https://github.com/UARK-AICV/MEGANet.

****

GazeGNN: A Gaze-Guided Graph Neural Network for Chest X-Ray Classification

Bin Wang, Hongyi Pan, Armstrong Aboah, Zheyuan Zhang, Elif Keles, Drew Torigian, Baris Turkbey, Elizabeth Krupinski, Jayaram Udupa, Ulas Bagci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2194-2203

Eye tracking research is important in computer vision because it can help us understand how humans interact with the visual world. Specifically for high-risk applications, such as in medical imaging, eye tracking can help us to comprehend how radiologists and other medical professionals search, analyze, and interpret images for diagnostic and clinical purposes. Hence, the application of eye tracking techniques in disease classification has become increasingly popular in recent years. Contemporary works usually transform gaze information collected by eye tracking devices into visual attention maps (VAMs) to supervise the learning process. However, this is a time-consuming preprocessing step, which stops us from applying eye tracking to radiologists' daily work. To solve this problem, we propose a novel gaze-guided graph neural network (GNN), GazeGNN, to leverage raw eye-gaze data without being converted into VAMs. In GazeGNN, to directly integrate eye gaze into image classification, we create a unified representation graph that models both images and gaze pattern information. With this benefit, we develop a real-time, real-world, end-to-end disease classification algorithm for the first time in the literature. This achievement demonstrates the practicality and feasibility of integrating real-time eye tracking techniques into the daily work of radiologists. To our best knowledge, GazeGNN is the first work that adopts GNN to integrate image and eye-gaze data. Our experiments on the public chest X-ray dataset show that our proposed method exhibits the best classification performance compared to existing methods. The code is available.

****

LipAT: Beyond Style Transfer for Controllable Neural Simulation of Lipstick Using Cosmetic Attributes

Amila Silva, Olga Moskvyak, Alexander Long, Ravi Garg, Stephen Gould, Gil Avraham, Anton van den Hengel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8046-8055

Lipstick virtual try-on (VTO) experiences have become widespread across the e-commerce sector and assist users in eliminating the guesswork of shopping online. However, such experiences still lack in both realism and accuracy. In this work, we propose LipAT, a neural framework that blends the strengths of Physics-Based Rendering (PBR) and Neural Style Transfer (NST) approaches to directly apply lipstick onto face images given lipstick attributes (e.g., colour, finish type). LipAT consists of a physics aware neural lipstick application module (LAM) to apply lipstick on face images given its attributes and Lipstick Refiner Module (LRM) to improve the realism by refining the imperfections. Unlike the NST approaches, LipAT allows precise and controllable lipstick attribute preservation, without requiring crude approximations and inference of various intertwined environment factors (e.g., scene lighting, face structure etc) involved in image generation that is required for accurate PBR. We propose an experimental framework with quantitative metrics to evaluate different desirable aspects of the lipstick attribute driven try-on alongside user studies to further validate our findings. Our results show that LipAT considerably outperforms fully-automated PBR approaches in preserving realism and the NST approaches in preserving various lipstick attributes such as finish types.

****

WildlifeDatasets: An Open-Source Toolkit for Animal Re-Identification

Vojt■ch ■ermák, Lukas Picek, Lukáš Adam, Kostas Papafitsoros; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5953-5963

In this paper, we present WildlifeDatasets - an open-source toolkit intended primarily for ecologists and computer-vision / machine-learning researchers. The WildlifeDatasets is written in Python, allows straightforward access to publicly available wildlife datasets, and provides a wide variety of methods for dataset p

re-processing, performance analysis, and model fine-tuning. We showcase the toolkit in various scenarios and baseline experiments, including, to the best of our knowledge, the most comprehensive experimental comparison of datasets and methods for wildlife re-identification, including both local descriptors and deep learning approaches. Furthermore, we provide the first-ever foundation model for individual re-identification within a wide range of species - MegaDescriptor - that provides state-of-the-art performance on animal re-identification datasets and outperforms other pre-trained models such as CLIP and DINOv2 by a significant margin. To make the model available to the general public and to allow easy integration with any existing wildlife monitoring applications, we provide multiple MegaDescriptor flavors (i.e., Small, Medium, and Large) through the HuggingFace hub.
****
OTAS: Unsupervised Boundary Detection for Object-Centric Temporal Action Segmentation
Yuerong Li, Zhengrong Xue, Huazhe Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6437-6446
Temporal action segmentation is typically achieved by discovering the dramatic variances in global visual descriptors. In this paper, we explore the merits of local features by proposing the unsupervised framework of Object-centric Temporal Action Segmentation (OTAS). Broadly speaking, OTAS consists of self-supervised global and local feature extraction modules as well as a boundary selection module that fuses the features and detects salient boundaries for action segmentation. As a second contribution, we discuss the pros and cons of existing frame-level and boundary-level evaluation metrics. Through extensive experiments, we find OTAS is superior to the previous state-of-the-art method by 41% on average in terms of our recommended F1 score. Surprisingly, OTAS even outperforms the ground-truth human annotations in the user study. Moreover, OTAS is efficient enough to allow real-time inference
****
Deblur-NSFF: Neural Scene Flow Fields for Blurry Dynamic Scenes
Achleshwar Luthra, Shiva Souhith Gantha, Xiyun Song, Heather Yu, Zongfang Lin, Liang Peng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3658-3667
In this work, we present a method to address the problem of novel view and time synthesis of complex dynamic scenes considering the input video is subject to blurriness caused due to camera or object motion or out-of-focus blur. Neural Scene Flow Field (NSFF) has shown remarkable results by training a dynamic NeRF to capture motion in the scene, but this method is not robust to unstable camera handling which can lead to blurred renderings. We propose Deblur-NSFF, a method that learns spatially-varying blur kernels to simulate the blurring process and gradually learns a sharp time-conditioned NeRF representation. We describe how to optimize our representation for sharp space-time view synthesis. Given blurry input frames, we perform both quantitative and qualitative comparison with state-of-the-art methods on modified NVIDIA Dynamic Scene dataset. We also compare our method with Deblur-NeRF, a method that has been designed to handle blur in static scenes. The demonstrated results show that our method outperforms prior work.
****
Multi-Source Domain Adaptation for Object Detection With Prototype-Based Mean Teacher
Atif Belal, Akhil Meethal, Francisco Perdigon Romero, Marco Pedersoli, Eric Granger; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1277-1286
Adapting visual object detectors to operational target domains is a challenging task, commonly achieved using unsupervised domain adaptation (UDA) methods. Recent studies have shown that when the labeled dataset comes from multiple source domains, treating them as separate domains and performing a multi-source domain adaptation (MSDA) improves the accuracy and robustness over blending these source domains and performing a UDA. For adaptation, existing MSDA methods learn domain-invariant and domain-specific parameters (for each source domain). However, un

like single-source UDA methods, learning domain-specific parameters makes them grow significantly in proportion to the number of source domains. This paper proposes a novel MSDA method called Prototype-based Mean Teacher (PMT), which uses class prototypes instead of domain-specific subnets to encode domain-specific information. These prototypes are learned using a contrastive loss, aligning the same categories across domains and separating different categories far apart. Given the use of prototypes, the number of parameters required for our PMT method does not increase significantly with the number of source domains, thus reducing memory issues and possible overfitting. Empirical studies indicate that PMT outperforms state-of-the-art MSDA methods on several challenging object detection datasets. Our code is available at https://github.com/imatif17/Prototype-Mean-Teacher

****

PathLDM: Text Conditioned Latent Diffusion Model for Histopathology

Srikar Yellapragada, Alexandros Graikos, Prateek Prasanna, Tahsin Kurc, Joel Saltz, Dimitris Samaras; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5182-5191

To achieve high-quality results, diffusion models must be trained on large datasets. This can be notably prohibitive for models in specialized domains, such as computational pathology. Conditioning on labeled data is known to help in data-efficient model training. Therefore, histopathology reports, which are rich in valuable clinical information, are an ideal choice as guidance for a histopathology generative model. In this paper, we introduce PathLDM, the first text-conditioned Latent Diffusion Model tailored for generating high-quality histopathology images. Leveraging the rich contextual information provided by pathology text reports, our approach fuses image and textual data to enhance the generation process. By utilizing GPT's capabilities to distill and summarize complex text reports, we establish an effective conditioning mechanism. Through strategic conditioning and necessary architectural enhancements, we achieved a SoTA FID score of 7.64 for text-to-image generation on the TCGA-BRCA dataset, significantly outperforming the closest text-conditioned competitor with FID 30.1.

****

EASUM: Enhancing Affective State Understanding Through Joint Sentiment and Emotion Modeling for Multimodal Tasks

Yewon Hwang, Jong-Hwan Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5668-5678

Multimodal sentiment analysis (MSA) and multimodal emotion recognition (MER) tasks have gained a surge of attention in recent years. Although both tasks share common ground in many ways, they are often treated as a separate task. In this work, we propose, EASUM, a new training scheme for bridging the MSA and MER tasks. EASUM aims to bring mutual benefits to both tasks based on the premise that the sentiment and emotion are closely related; hence each information should provide deeper insight into one's affective state to complement the other. We exploit this premise to further improve the performance of each task by 1) first training a domain general model using four benchmark datasets from the MSA and MER tasks: CMU-MOSI, CMU-MOSEI, MELD, and IEMOCAP. Depending on the dataset, the domain general model learns to predict sentiment or emotion values based on the domain invariant features. 2) Then these values are later used as auxiliary pseudo labels when training a domain specific model for each task. Our premise as well as new training scheme are validated through extensive experiments on the four benchmark datasets. The results also demonstrate that the proposed method outperforms the state-of-the-art on the CMU-MOSI, CMU-MOSEI, and MELD datasets, and performs comparable to the state-of-the-art on the IEMOCAP dataset while using approximately 40% fewer parameters.

****

Efficient Explainable Face Verification Based on Similarity Score Argument Backpropagation

Marco Huber, Anh Thi Luu, Philipp Terhörst, Naser Damer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4736-4745

Explainable Face Recognition is gaining growing attention as the use of the technology is gaining ground in security-critical applications. Understanding why two face images are matched or not matched by a given face recognition system is important to operators, users, and developers to increase trust, accountability, develop better systems, and highlight unfair behavior. In this work, we propose a similarity score argument backpropagation (xSSAB) approach that supports or opposes the face-matching decision to visualize spatial maps that indicate similar and dissimilar areas as interpreted by the underlying FR model. Furthermore, we present Patch-LFW, a new explainable face verification benchmark that enables along with a novel evaluation protocol, the first quantitative evaluation of the validity of similarity and dissimilarity maps in explainable face recognition approaches. We compare our efficient approach to state-of-the-art approaches demonstrating a superior trade-off between efficiency and performance. The code as well as the proposed Patch-LFW is publicly available at: https://github.com/marcohuber/xSSAB.

****

Rank2Tell: A Multimodal Driving Dataset for Joint Importance Ranking and Reasoning

Enna Sachdeva, Nakul Agarwal, Suhas Chundi, Sean Roelofs, Jiachen Li, Mykel Kochenderfer, Chiho Choi, Behzad Dariush; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7513-7522

The widespread adoption of commercial autonomous vehicles (AVs) and advanced driver assistance systems (ADAS) may largely depend on their acceptance by society, for which their perceived trustworthiness and interpretability to riders are crucial. In general, this task is challenging because modern autonomous systems software relies heavily on black-box artificial intelligence models. Towards this goal, this paper introduces a novel dataset, Rank2Tell, a multi-modal ego-centric dataset for Ranking the importance level and Telling the reason for the importance. Using various close and open-ended visual question answering, the dataset provides dense annotations of various semantic, spatial, temporal, and relational attributes of various important objects in complex traffic scenarios. The dense annotations and unique attributes of the dataset make it a valuable resource for researchers working on visual scene understanding and related fields. Furthermore, we introduce a joint model for joint importance level ranking and natural language captions generation to benchmark our dataset and demonstrate performance with quantitative evaluations.

****

ArcAid: Analysis of Archaeological Artifacts Using Drawings

Offry Hayon, Stefan Münger, Ilan Shimshoni, Ayellet Tal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7264-7274

Archaeology is an intriguing domain for computer vision. It suffers not only from shortage in (labeled) data, but also from highly-challenging data, which is often extremely abraded and damaged. This paper proposes a novel semi-supervised model for classification and retrieval of images of archaeological artifacts. This model utilizes unique data that exists in the domain--manual drawings made by special artists. These are used during training to implicitly transfer the domain knowledge from the drawings to their corresponding images, improving their classification results. We show that while learning how to classify, our model also learns how to generate drawings of the artifacts, an important documentation task, which is currently performed manually. Last but not least, we collected a new dataset of stamp-seals of the Southern Levant. Our code and dataset are publicly available.

****

FishTrack23: An Ensemble Underwater Dataset for Multi-Object Tracking

Matthew Dawkins, Jack Prior, Bryon Lewis, Robin Faillettaz, Thompson Banez, Mary Salvi, Audrey Rollo, Julien Simon, Matthew Campbell, Matthew Lucero, Aashish Chaudhary, Benjamin Richards, Anthony Hoogs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7167-7176

Tracking and classifying fish in optical underwater imagery presents several cha llenges which are encountered less frequently in terrestrial domains. Video may contain large schools comprised of many individuals, dynamic natural backgrounds, highly variable target scales, volatile collection conditions, and non-fish moving confusers including debris, marine snow, and other organisms. Additionally, there is a lack of large public datasets for algorithm evaluation available in this domain. The contributions of this paper is three fold. First, we present the FishTrack23 dataset which provides a large quantity of expert-annotated fish groundtruth tracks, in imagery and video collected across a range of different backgrounds, locations, collection conditions, and organizations. Approximately 850k bounding boxes across 26k tracks are included in the release of the ensemble, with potential for future growth in later releases. Second, we evaluate improvements upon baseline object detectors, trackers and classifiers on the dataset. Lastly, we integrate these methods into web and desktop interfaces to expedite annotation generation on new datasets.

****

Reducing the Side-Effects of Oscillations in Training of Quantized YOLO Networks

Kartik Gupta, Akshay Asthana; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2452-2461

Quantized networks use less computational and memory resources and are suitable for deployment on edge devices. While quantization-aware training QAT is the well-studied approach to quantize the networks at low precision, most research focuses on over-parameterized networks for classification with limited studies on popular and edge device friendly single-shot object detection and semantic segmentation methods like YOLO. Moreover, majority of QAT methods rely on Straight-through Estimator (STE) approximation which suffers from an oscillation phenomenon resulting in sub-optimal network quantization. In this paper, we show that it is difficult to achieve extremely low precision (4-bit and lower) for efficient YOLO models even with SOTA QAT methods due to oscillation issue and existing methods to overcome this problem are not effective on these models. To mitigate the effect of oscillation, we first propose Exponentially Moving Average (EMA) based update to the QAT model. Further, we propose a simple QAT correction method, namely QC, that takes only a single epoch of training after standard QAT procedure to correct the error induced by oscillating weights and activations resulting in a more accurate quantized model. With extensive evaluation on COCO dataset using various YOLO5 and YOLO7 variants, we show that our correction method improves quantized YOLO networks consistently on both object detection and segmentation tasks at low-precision (4-bit and 3-bit).

****

PressureVision++: Estimating Fingertip Pressure From Diverse RGB Images

Patrick Grady, Jeremy A. Collins, Chengcheng Tang, Christopher D. Twigg, Kunal Aneja, James Hays, Charles C. Kemp; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8698-8708

Touch plays a fundamental role in manipulation for humans; however, machine perception of contact and pressure typically requires invasive sensors. Recent research has shown that deep models can estimate hand pressure based on a single RGB image. However, evaluations have been limited to controlled settings since collecting diverse data with ground-truth pressure measurements is difficult. We present a novel approach that enables diverse data to be captured with only an RGB camera and a cooperative participant. Our key insight is that people can be prompted to apply pressure in a certain way, and this prompt can serve as a weak label to supervise models to perform well under varied conditions. We collect a novel dataset with 51 participants making fingertip contact with diverse objects. Our network, PressureVision++, outperforms human annotators and prior work. We also demonstrate an application of PressureVision++ to mixed reality where pressure estimation allows everyday surfaces to be used as arbitrary touch-sensitive interfaces. Code, data, and models are available online.

****

Diffusion Models Meet Image Counter-Forensics

Matías Tailanián, Marina Gardella, Alvaro Pardo, Pablo Musé; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp.

3925-3935

From its acquisition in the camera sensors to its storage, different operations are performed to generate the final image. This pipeline imprints specific traces into the image to form a natural watermark. Tampering with an image disturbs these traces; these disruptions are clues that are used by most methods to detect and locate forgeries. In this article, we assess the capabilities of diffusion models to erase the traces left by forgers and, therefore, deceive forensics methods. Such an approach has been recently introduced for adversarial purification, achieving significant performance. We show that diffusion purification methods are well suited for counter-forensics tasks. Such approaches outperform already existing counter-forensics techniques both in deceiving forensics methods, and in preserving the natural look of the purified images. The source code will be provided upon acceptance.

****

STYLIP: Multi-Scale Style-Conditioned Prompt Learning for CLIP-Based Domain Generalization

Shirsha Bose, Ankit Jha, Enrico Fini, Mainak Singha, Elisa Ricci, Biplab Banerjee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5542-5552

arge-scale foundation models, such as CLIP, have demonstrated impressive zero-shot generalization performance on downstream tasks, leveraging well-designed language prompts. However, these prompt learning techniques often struggle with domain shift, limiting their generalization capabilities. In our study, we tackle this issue by proposing STYLIP, a novel approach for Domain Generalization (DG) that enhances CLIP's classification performance across domains. Our method focuses on a domain-agnostic prompt learning strategy, aiming to disentangle the visual style and content information embedded in CLIP's pre-trained vision encoder, enabling effortless adaptation to novel domains during inference. To achieve this, we introduce a set of style projectors that directly learn the domain-specific prompt tokens from the extracted multi-scale style features. These generated prompt embeddings are subsequently combined with the multi-scale visual content features learned by a content projector. The projectors are trained in a contrastive manner, utilizing CLIP's fixed vision and text backbones. Through extensive experiments conducted in five different DG settings on multiple benchmark datasets, we consistently demonstrate that STYLIP outperforms the current state-of-the-art (SOTA) methods.

****

Increasing Biases Can Be More Efficient Than Increasing Weights

Carlo Metta, Marco Fantozzi, Andrea Papini, Gianluca Amato, Matteo Bergamaschi, Silvia Giulia Galfrè, Alessandro Marchetti, Michelangelo Vegliò, Maurizio Parton, Francesco Morandin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2810-2819

We introduce a novel computational unit for neural networks that features multiple biases, challenging the traditional perceptron structure. This unit emphasizes the importance of preserving uncorrupted information as it is passed from one unit to the next, applying activation functions later in the process with specialized biases for each unit. Through both empirical and theoretical analyses, we show that by focusing on increasing biases rather than weights, there is potential for significant enhancement in a neural network model's performance. This approach offers an alternative perspective on optimizing information flow within neural networks. Commented source code at https://github.com/CuriosAI/dac-dev.

****

TransRadar: Adaptive-Directional Transformer for Real-Time Multi-View Radar Semantic Segmentation

Yahia Dalbah, Jean Lahoud, Hisham Cholakkal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 353-362

Scene understanding plays an essential role in enabling autonomous driving and maintaining high standards of performance and safety. To address this task, cameras and laser scanners (LiDARs) have been the most commonly used sensors, with radars being less popular. Despite that, radars remain low-cost, information-dense

, and fast-sensing techniques that are resistant to adverse weather conditions. While multiple works have been previously presented for radar-based scene semantic segmentation, the nature of the radar data still poses a challenge due to the inherent noise and sparsity, as well as the disproportionate foreground and background. In this work, we propose a novel approach to the semantic segmentation of radar scenes using a multi-input fusion of radar data through a novel architecture and loss functions that are tailored to tackle the drawbacks of radar perception. Our novel architecture includes an efficient attention block that adaptively captures important feature information. Our method, TransRadar, outperforms state-of-the-art methods on the CARRADA and RADIal datasets while having smaller model sizes.

****

Sequential Transformer for End-to-End Video Text Detection

Jun-Bo Zhang, Meng-Biao Zhao, Fei Yin, Cheng-Lin Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6520-6530

In existing methods of video text detection, the detection and tracking branches are usually independent of each other, and although they jointly optimize the backbone network, the tracking-by-detection paradigm still needs to be used during the inference stage. To address this issue, we propose a novel video text detection framework based on sequential transformer, which decodes detection and tracking tasks in parallel, without explicitly setting up a tracking branch. To achieve this, we first introduce the concept of instance query, which learns long-term context information in the video sequence. Then, based on the instance query, the transformer decoder is used to predict the entire box and mask sequence of the text instance in one pass. As a result, the tracking task is realized naturally. In addition, the proposed method can be applied to the scene text detection task seamlessly, without modifying any modules. To the best of our knowledge, this is the first framework to unify the tasks of scene text detection and video text detection. Our model achieves state-of-the-art performance on four video text datasets (YVT, RT-1K, BOVText, and BiRViT-1K), and competitive results on three scene text datasets (CTW1500, MSRA-TD500, and Total-Text). The code is available at https://github.com/zjb-1/SeqVideoText.

****

The Background Also Matters: Background-Aware Motion-Guided Objects Discovery

Sandra Kara, Hejer Ammar, Florian Chabot, Quoc-Cuong Pham; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1216-1225

Recent works have shown that objects discovery can largely benefit from the inherent motion information in video data. However, these methods lack a proper background processing, resulting in an over-segmentation of the non-object regions into random segments. This is a critical limitation given the unsupervised setting, where object segments and noise are not distinguishable. To address this limitation we propose BMOD, a Background-aware Motion-guided Objects Discovery method. Concretely, we leverage masks of moving objects extracted from optical flow and design a learning mechanism to extend them to the true foreground composed of both moving and static objects. The background, a complementary concept of the learned foreground class, is then isolated in the object discovery process. This enables a joint learning of the objects discovery task and the object/non-object separation. The conducted experiments on synthetic and real-world datasets show that integrating our background handling with various cutting-edge methods brings each time a considerable improvement. Specifically, we improve the objects discovery performance with a large margin, while establishing a strong baseline for object/non-object separation.

****

Neural Style Protection: Counteracting Unauthorized Neural Style Transfer

Yaxin Li, Jie Ren, Han Xu, Hui Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3966-3975

Arbitrary neural style transfer is an advanced AI technique that can effectively synthesize pictures with an artistic style similar to a given source picture. H owever, if such an AI technique is leveraged by unauthorized individuals, it can significantly infringe upon the copyright of the source picture's owner. In this paper, we study how to protect the artistic style of source images against unauthorized style transfer by adding imperceptible perturbations to the original source pictures. In particular, our goal is to disable the neural style transfer models from producing high-quality pictures with a similar style to the source pictures with slight manipulating the source images. We introduce Neural Style Protection (NSP), which provides protection for source images against various neural style transfer models. Through extensive experiments, we demonstrate the effectiveness and generalizability of the proposed style protection algorithm across numerous style transfer models using varied metrics.

****

FRoG-MOT: Fast and Robust Generic Multiple-Object Tracking by IoU and Motion-State Associations

Takuya Ogawa, Takashi Shibata, Toshinori Hosoi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6563-6572

This paper proposes a generic multi-object tracking (MOT) algorithm that is robust to unexpected motion changes for generic objects. Deep learning has dramatically been improving MOT performances. Nevertheless, state-of-the-art tracking algorithms are still sensitive to unexpected motion changes and the generic object target beyond person tracking. This is because standard MOT benchmark datasets such as MOT17 mainly consist of persons in a crowd, often lacking unexpected shape and motion changes; thus, these issues have yet to be focused on. We propose a simple-yet-effective MOT framework that can dynamically improve tracking continuity by associating each target based on adaptively modified motion states. The keys are 1) to represent the target motions using multiple motion states that have weak correlations with each other and 2) to modify those states that have the lowest similarity to past states as outliers. Our approach can overwhelmingly improve trajectory continuity and robustness to unexpected motion changes for generic objects. Comprehensive experiments have confirmed that our framework is comparable to existing state-of-the-art methods on a standard dataset and outperforms those algorithms on the GMOT dataset with an overall 2% improvement in IDF1, a measure of tracking continuity.

****

OVeNet: Offset Vector Network for Semantic Segmentation

Stamatis Alexandropoulos, Christos Sakaridis, Petros Maragos; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7407-7418

Semantic segmentation is a fundamental task in visual scene understanding. We focus on the supervised setting, where ground-truth semantic annotations are available. Based on knowledge about the high regularity of real-world scenes, we propose a method for improving class predictions by learning to selectively exploit information from neighboring pixels. In particular, our method is based on the prior that for each pixel, there is a seed pixel in its close neighborhood sharing the same prediction with the former. Motivated by this prior, we design a novel two-head network, named Offset Vector Network (OVeNet), which generates both standard semantic predictions and a dense 2D offset vector field indicating the offset from each pixel to the respective seed pixel, which is used to compute an alternative, seed-based semantic prediction. The two predictions are adaptively fused at each pixel using a learnt dense confidence map for the predicted offset vector field. We supervise offset vectors indirectly via optimizing the seed-based prediction and via a novel loss on the confidence map. Compared to the baseline state-of-the-art architectures HRNet and HRNet+OCR on which OVeNet is built, the latter achieves significant performance gains on three prominent benchmarks for semantic segmentation, namely Cityscapes, ACDC and ADE20K. Code is available at https://github.com/stamatisalex/OVeNet.

****

A Neural Height-Map Approach for the Binocular Photometric Stereo Problem

Fotios Logothetis, Ignas Budvytis, Roberto Cipolla; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1568-1577

In this work we propose a novel, highly practical, binocular photometric stereo (PS) framework, which has same acquisition speed as single view PS, however significantly improves the quality of the estimated geometry. As in recent neural multi-view shape estimation frameworks such as NeRF, SIREN and inverse graphics approach to multi-view photometric stereo (e.g. PS-NeRF) we formulate shape estimation task as learning of a differentiable surface and texture representation by minimising surface normal discrepancy for normals estimated from multiple varying light images for two views as well as discrepancy between rendered surface intensity and observed images. Our method differs from typical multi-view shape estimation approaches in two key ways. First, our surface is represented not as a volume but as a neural heightmap where heights of points on a surface are computed by a deep neural network. Second, instead of predicting an average intensity as PS-NeRF or introducing lambertian material assumptions as Guo et al., we use a learnt BRDF and perform near-field per point intensity rendering. Our method achieves the state-of-the-art performance on the DiLiGenT-MV dataset adapted to binocular stereo setup as well as a new binocular photometric stereo dataset - LUCES-ST.

****

Towards Addressing the Misalignment of Object Proposal Evaluation for Vision-Language Tasks via Semantic Grounding

Joshua Feinglass, Yezhou Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4397-4407

Object proposal generation serves as a standard pre-processing step in Vision-Language (VL) tasks (image captioning, visual question answering, etc.). The performance of object proposals generated for VL tasks is currently evaluated across all available annotations, a protocol that we show is misaligned - higher scores do not necessarily correspond to improved performance on downstream VL tasks. Our work serves as a study of this phenomenon and explores the effectiveness of semantic grounding to mitigate its effects. To this end, we propose evaluating object proposals against only a subset of available annotations, selected by thresholding an annotation importance score. Importance of object annotations to VL tasks is quantified by extracting relevant semantic information from text describing the image. We show that our method is consistent and demonstrates greatly improved alignment with annotations selected by image captioning metrics and human annotation when compared against existing techniques. Lastly, we compare current detectors used in the Scene Graph Generation (SGG) benchmark as a use case, which serves as an example of when traditional object proposal evaluation techniques are misaligned.

****

Spiking Neural Networks for Active Time-Resolved SPAD Imaging

Yang Lin, Edoardo Charbon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8147-8156

Single-photon avalanche diodes (SPADs) are detectors capable of capturing single photons and of performing photon counting. SPADs have an exceptional temporal resolution and are thus highly suitable for time-resolved imaging applications. Applications span from biomedical research to consumers with SPADs integrated in smartphones and mixed-reality headsets. While conventional SPAD imaging systems typically employ photon time-tagging and histogram-building in the workflow, the pulse signal output of a SPAD naturally lends itself as input to spiking neural networks (SNNs). Leveraging this potential, SNNs offer real-time, energy-efficient, and intelligent processing with high throughput. In this paper, we propose two SNN frameworks, namely the Transporter SNN and the Reversed Start-stop SNN, along with corresponding hardware schemes for active time-resolved SPAD imaging. These frameworks convert phase-coded spike trains into density- and interspike-interval-coded ones, enabling training with rate-based warm-up and Surrogate Gradient. The SNNs are evaluated on fluorescence lifetime imaging. The results demonstrate that the accuracy of shallow SNNs is on par with established benchmarks. Our vision is to integrate SNNs in SPAD sensors and to explore advanced SNNs within the proposed schemes for high-level applications.

****

Domain Generalization With Correlated Style Uncertainty
Zheyuan Zhang, Bin Wang, Debesh Jha, Ugur Demir, Ulas Bagci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2000-2009

Domain generalization (DG) approaches intend to extract domain invariant features that can lead to a more robust deep learning model. In this regard, style augmentation is a strong DG method taking advantage of instance-specific feature statistics containing informative style characteristics to synthetic novel domains. While it is one of the state-of-the-art methods, prior works on style augmentation have either disregarded the interdependence amongst distinct feature channels or have solely constrained style augmentation to linear interpolation. To address these research gaps, in this work, we introduce a novel augmentation approach, named Correlated Style Uncertainty (CSU), surpassing the limitations of linear interpolation in style statistic space and simultaneously preserving vital correlation information. Our method's efficacy is established through extensive experimentation on diverse cross-domain computer vision and medical imaging classification tasks: PACS, Office-Home, and Camelyon17 datasets, and the Duke-Market1501 instance retrieval task. The results showcase a remarkable improvement margin over existing state-of-the-art techniques. The source code is available: https://github.com/freshman97/CSU.

****

Leveraging Next-Active Objects for Context-Aware Anticipation in Egocentric Videos
Sanket Thakur, Cigdem Beyan, Pietro Morerio, Vittorio Murino, Alessio Del Bue; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8657-8666

Objects are crucial for understanding human-object interactions. By identifying the relevant objects, one can also predict potential future interactions or actions that may occur with these objects. In this paper, we study the problem of Short-Term Object interaction anticipation (STA) and propose NAOGAT (Next-Active-Object Guided Anticipation Transformer), a multi-modal end-to-end transformer network, that attends to objects in observed frames in order to anticipate the next-active-object (NAO) and, eventually, to guide the model to predict context-aware future actions. The task is challenging since it requires anticipating future action along with the object with which the action occurs and the time after which the interaction will begin, a.k.a. the time to contact (TTC). Compared to existing video modeling architectures for action anticipation, NAOGAT captures the relationship between objects and the global scene context in order to predict detections for the next active object and anticipate relevant future actions given these detections, leveraging the objects' dynamics to improve accuracy. One of the key strengths of our approach, in fact, is its ability to exploit the motion dynamics of objects within a given clip, which is often ignored by other models, and separately decoding the object-centric and motion-centric information. Through our experiments, we show that our model outperforms existing methods on two separate datasets, Ego4D and EpicKitchens-100 ("Unseen Set"), as measured by several additional metrics, such as time to contact, and next-active-object localization.

****

CryoRL: Reinforcement Learning Enables Efficient Cryo-EM Data Collection
Quanfu Fan, Yilai Li, Yuguang Yao, John Cohn, Sijia Liu, Ziping Xu, Seychelle Vos, Michael Cianfrocco; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7892-7902

Single-particle cryo-electron microscopy (cryo-EM) has become one of the mainstream structural biology techniques because of its ability to determine high-resolution structures of dynamic bio-molecules. However, cryo-EM data acquisition remains expensive and labor-intensive, requiring substantial expertise. Structural biologists need a more efficient and objective method to collect the best data in a limited time frame. We formulate the cryo-EM data collection task as an optimization problem in this work. The goal is to maximize the total number of good images taken within a specified period. We show that reinforcement learning offe

rs an effective way to plan cryo-EM data collection, successfully navigating heterogenous cryo-EM grids. The approach we developed, cryoRL, demonstrates better performance than average users for data collection under similar settings.
****

On the Fly Neural Style Smoothing for Risk-Averse Domain Generalization

Akshay Mehra, Yunbei Zhang, Bhavya Kailkhura, Jihun Hamm; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3800-3811

Achieving high accuracy on data from domains unseen during training is a fundamental challenge in domain generalization (DG). While state-of-the-art DG classifiers have demonstrated impressive performance across various tasks, they have shown a bias towards domain-dependent information, such as image styles, rather than domain-invariant information, such as image content. This bias renders them unreliable for deployment in risk-sensitive scenarios such as autonomous driving where a misclassification could lead to catastrophic consequences. To enable risk-averse predictions from a DG classifier, we propose a novel inference procedure, Test-Time Neural Style Smoothing (TT-NSS), that uses a "style-smoothed" version of the DG classifier for prediction at test time. Specifically, the style-smoothed classifier classifies a test image as the most probable class predicted by the DG classifier on random re-stylizations of the test image. TT-NSS uses a neural style transfer module to stylize a test image on the fly, requires only black-box access to the DG classifier, and crucially, abstains when predictions of the DG classifier on the stylized test images lack consensus. Additionally, we propose a neural style smoothing (NSS) based training procedure that can be seamlessly integrated with existing DG methods. This procedure enhances prediction consistency, improving the performance of TT-NSS on non-abstained samples. Our empirical results demonstrate the effectiveness of TT-NSS and NSS at producing and improving risk-averse predictions on unseen domains from DG classifiers trained with SOTA training methods on various benchmark datasets and their variations.
****

StyleGenes: Discrete and Efficient Latent Distributions for GANs

Evangelos Ntavelis, Mohamad Shahbazi, Iason Kastanis, Martin Danelljan, Luc Van Gool; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4077-4086

We propose a discrete latent distribution for Generative Adversarial Networks (GANs). Instead of drawing latent vectors from a continuous prior, we sample from a finite set of learnable latents. However, a direct parametrization of such a distribution leads to an intractable linear increase in memory in order to ensure sufficient sample diversity. We address this key issue by taking inspiration from the encoding of information in biological organisms. Instead of learning a separate latent vector for each sample, we split the latent space into a set of genes. For each gene, we train a small bank of gene variants. Thus, by independently sampling a variant for each gene and combining them into the final latent vector, our approach can represent a vast number of unique latent samples from a compact set of learnable parameters. Interestingly, our gene-inspired latent encoding allows for new and intuitive approaches to latent-space exploration, enabling conditional sampling from our unconditionally trained model. Our approach preserves state-of-the-art photo-realism while achieving better disentanglement than the widely-used StyleMapping network.
****

Aligning Non-Causal Factors for Transformer-Based Source-Free Domain Adaptation

Sunandini Sanyal, Ashish Ramayee Asokan, Suvaansh Bhambri, Pradyumna YM, Akshay Kulkarni, Jogendra Nath Kundu, R. Venkatesh Babu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1904-1913

Conventional domain adaptation algorithms aim to achieve better generalization by aligning only the task-discriminative causal factors between a source and target domain. However, we find that retaining the spurious correlation between causal and non-causal factors plays a vital role in bridging the domain gap and improving target adaptation. Therefore, we propose to build a framework that disentangles and supports causal factor alignment by aligning the non-causal factors fi rst. We also investigate and find that the strong shape bias of vision transformers, coupled with its multi-head attentions, make it a suitable architecture for realizing our proposed disentanglement. Hence, we propose to build a Causality-enforcing Source Free Transformer framework (C-SFTrans) to achieve dis entanglement via a novel two-stage alignment approach: a) non-causal factor alignment: non-causal factors are aligned using a style classification task which leads to an overall global alignment, b) task-discriminative causal factor alignment: causal factors are aligned via target adaptation. We are the first to investigate the role of vision transformers (ViTs) in a privacy-preserving source-free setting. Our approach achieves state-of-the-art results in several DA benchmarks.

****

Benchmarking Out-of-Distribution Detection in Visual Question Answering

Xiangxi Shi, Stefan Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5485-5495

When faced with an out-of-distribution (OOD) question or image, visual question answering (VQA) systems may provide unreliable answers. If relied on by real users or secondary systems, these failures may range from annoying to potentially endangering. Detecting OOD samples in single-modality settings is well-studied; however, limited attention has been paid to vision-and-language settings. In this work, we examine the question of OOD detection in the multimodal VQA task and benchmark a suite of approaches to identify OOD image-question pairs. In our experiments, we leverage popular VQA datasets to benchmark detection performance across a range of difficulties. We also produce composite datasets to examine impacts of individual modalities and of image-question agreement. Our results show that answer confidence alone is often a poor signal and that methods based on image-based question generation or examining model attention can lead to significantly better results. We find detecting ungrounded image-question pairs and small shifts in image distribution remain challenging.

****

Joint Depth Prediction and Semantic Segmentation With Multi-View SAM

Mykhailo Shvets, Dongxu Zhao, Marc Niethammer, Roni Sengupta, Alexander C. Berg; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1328-1338

Multi-task approaches to joint depth and segmentation prediction are well-studied for monocular images. Yet, predictions from a single-view are inherently limited, while multiple views are available in many robotics applications. On the other end of the spectrum, video-based and full 3D methods require numerous frames to perform reconstruction and segmentation. With this work we propose a Multi-View Stereo (MVS) technique for depth prediction that benefits from rich semantic features of the Segment Anything Model (SAM). This enhanced depth prediction, in turn, serves as a prompt to our Transformer-based semantic segmentation decoder. We report the mutual benefit that both tasks enjoy in our quantitative and qualitative studies on the ScanNet dataset. Our approach consistently outperforms single-task MVS and segmentation models, along with multi-task monocular methods.

****

GC-VTON: Predicting Globally Consistent and Occlusion Aware Local Flows With Neighborhood Integrity Preservation for Virtual Try-On

Hamza Rawal, Muhammad Junaid Ahmad, Farooq Zaman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5251-5260

Flow based garment warping is an integral part of image-based virtual try-on networks. However, optimizing a single flow predicting network for simultaneous global boundary alignment and local texture preservation results in sub-optimal flow fields. Moreover, dense flows are inherently not suited to handle intricate conditions like garment occlusion by body parts or by other garments. Forcing flows to handle the above issues results in various distortions like texture squeezing, and stretching. In this work, we propose a novel approach where we disentangle the global boundary alignment and local texture preserving tasks via our GlobalNet and LocalNet modules. A consistency loss is then employed between the two modules which harmonizes the local flows with the global boundary alignment. Additionally, we explicitly handle occlusions by predicting body-parts visibility m ask, which is used to mask out the occluded regions in the warped garment. The masking prevents the LocalNet from predicting flows that distort texture to compensate for occlusions. We also introduce a novel regularization loss (NIPR), that defines a criteria to identify the regions in the warped garment where texture integrity is violated (squeezed or stretched). NIPR subsequently penalizes the flow in those regions to ensure regular and coherent warps that preserve the texture in local neighborhoods. Evaluation on a widely used virtual try-on dataset demonstrates strong performance of our network compared to the current SOTA methods.

****

Enforcing Sparsity on Latent Space for Robust and Explainable Representations

Hanao Li, Tian Han; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5282-5291

Recently, dense latent variable models have shown promising results, but their distributed and potentially redundant codes make them less interpretable and less robust to noise. On the other hand, sparse representations are more parsimonious, providing better explainability and noise robustness, but it is difficult to enforce sparsity due to the complexity and computational cost involved. In this paper, we propose a novel unsupervised learning approach to enforce sparsity on the latent space for the generator model, utilizing a gradually sparsified spike and slab distribution as our prior. Our model is composed of a top-down generator network that maps the latent variable to the observations. We use maximum likelihood sampling to infer latent variables in the generator's posterior direction, and spike and slab regularization in the inference stage can induce sparsity by pushing non-informative latent dimensions toward zero. Our experiments show that the learned sparse latent representations preserve the majority of the information, and our model can learn disentangled semantics, increase the explainability of the latent codes, and enhance the robustness of the classification and denoising tasks.

****

Unsupervised Domain Adaptation for Semantic Segmentation With Pseudo Label Self-Refinement

Xingchen Zhao, Niluthpol Chowdhury Mithun, Abhinav Rajvanshi, Han-Pang Chiu, Supun Samarasekera; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2399-2409

Deep learning-based solutions for semantic segmentation suffer from significant performance degradation when tested on data with different characteristics than what was used during the training. Adapting the models using annotated data from the new domain is not always practical. Unsupervised Domain Adaptation (UDA) approaches are crucial in deploying these models in the actual operating conditions. Recent state-of-the-art (SOTA) UDA methods employ a teacher-student self-training approach, where a teacher model is used to generate pseudo-labels for the new data which in turn guide the training process of the student model. Though this approach has seen a lot of success, it suffers from the issue of noisy pseudo-labels being propagated in the training process. To address this issue, we propose an auxiliary pseudo-label refinement network (PRN) for online refining of the pseudo labels and also localizing the pixels whose predicted labels are likely to be noisy. Being able to improve the quality of pseudo labels and select highly reliable ones, PRN helps self-training of segmentation models to be robust against pseudo label noise propagation during different stages of adaptation. We evaluate our approach on benchmark datasets with three different domain shifts, and our approach consistently performs significantly better than the previous state-of-the-art methods.

****

HalluciDet: Hallucinating RGB Modality for Person Detection Through Privileged Information

Heitor Rapela Medeiros, Fidel A. Guerrero Peña, Masih Aminbeidokhti, Thomas Dubail, Eric Granger, Marco Pedersoli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1444-1453

A powerful way to adapt a visual recognition model to a new domain is through im age translation. However, common image translation approaches only focus on generating data from the same distribution as the target domain. Given a cross-modal application, such as pedestrian detection from aerial images, with a considerable shift in data distribution between infrared (IR) to visible (RGB) images, a translation focused on generation might lead to poor performance as the loss focuses on irrelevant details for the task. In this paper, we propose HalluciDet, an IR-RGB image translation model for object detection. Instead of focusing on reconstructing the original image on the IR modality, it seeks to reduce the detection loss of an RGB detector, and therefore avoids the need to access RGB data. This model produces a new image representation that enhances objects of interest in the scene and greatly improves detection performance. We empirically compare our approach against state-of-the-art methods for image translation and for fine-tuning on IR, and show that our HalluciDet improves detection accuracy in most cases by exploiting the privileged information encoded in a pre-trained RGB detector. Code: https://github.com/heitorrapela/HalluciDet.

****

Improving Fairness in Deepfake Detection

Yan Ju, Shu Hu, Shan Jia, George H. Chen, Siwei Lyu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4655-4665

Despite the development of effective deepfake detectors in recent years, recent studies have demonstrated that biases in the data used to train these detectors can lead to disparities in detection accuracy across different races and genders. This can result in different groups being unfairly targeted or excluded from detection, allowing undetected deepfakes to manipulate public opinion and erode trust in a deepfake detection model. While existing studies have focused on evaluating fairness of deepfake detectors, to the best of our knowledge, no method has been developed to encourage fairness in deepfake detection at the algorithm level. In this work, we make the first attempt to improve deepfake detection fairness by proposing novel loss functions that handle both the setting where demographic information (e.g., annotations of race and gender) is available as well as the case where this information is absent. Fundamentally, both approaches can be used to convert many existing deepfake detectors into ones that encourages fairness. Extensive experiments on four deepfake datasets and five deepfake detectors demonstrate the effectiveness and flexibility of our approach in improving deepfake detection fairness. Our code is available at https://github.com/littlejuyan/DF_Fairness.

****

Evolve: Enhancing Unsupervised Continual Learning With Multiple Experts

Xiaofan Yu, Tajana Rosing, Yunhui Guo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2366-2377

Recent years have seen significant progress in unsupervised continual learning methods. Despite their success in controlled settings, their practicality in real-world contexts remains uncertain. In this paper, we first empirically investigate existing self-supervised continual learning methods. We show that even with a replay buffer, existing methods cannot preserve the critical knowledge on videos with temporal-correlated input. Our insight is that the primary challenge of unsupervised continual learning stems from the unpredictable input and the absence of supervision as well as prior knowledge. Drawing inspiration from hybrid AI, we introduce EVOLVE, an innovative framework employing multiple pre-trained models in the cloud, as experts, to bolster existing self-supervised learning methods on local clients. EVOLVE harnesses expert guidance through a novel expert aggregation loss, calculated and returned from the cloud. It also dynamically assigns weights to experts based on their confidence and tailored prior knowledge, thereby offering adaptive supervision for new streaming data. We extensively validate EVOLVE across several real-world data streams with temporal correlation. The results convincingly demonstrate that EVOLVE surpasses the best state-of-the-art unsupervised continual learning method by 6.1-53.7% in top-1 linear evaluation accuracy across various data streams, affirming the efficacy of diverse expert guidance. The codebase is at https://github.com/Orienfish/Evolve.

****

NeRFEditor: Differentiable Style Decomposition for 3D Scene Editing

Chunyi Sun, Yanbin Liu, Junlin Han, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7306-7315

We present NeRFEditor, an efficient learning framework for 3D scene editing, which takes a video as input and outputs a high quality, identity-preserving stylized 3D scene. Our goal is to bridge the gap between 2D and 3D editing, catering to a wide array of creative modifications such as reference-guided alterations, text-based prompts, and user interactions. We achieve this by encouraging a pre-trained StyleGAN model and a NeRF model to learn mutually consistent renderings. Specifically, we use NeRF to generate numerous (image, camera pose)-pairs to train an adjustor module, which adapts the StyleGAN latent code for generating high fidelity stylized images from any given viewing angle. To extrapolate edits to novel views, i.e., those not seen by StyleGAN pre-training, while maintaining 360 degree consistency, we propose a second self-supervised module that maps these views into the hidden space of StyleGAN. Together these two modules produce sufficient guidance for NeRF to learn consistent stylization effects across the full range of views. Experiments show that NeRFEditor outperforms prior work on benchmark and real-world scenes with better editability, fidelity, and identity preservation.

****

Personalized Face Inpainting With Diffusion Models by Parallel Visual Attention

Jianjin Xu, Saman Motamed, Praneetha Vaddamanu, Chen Henry Wu, Christian Haene, Jean-Charles Bazin, Fernando De la Torre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5432-5442

Face inpainting is important in various applications, such as photo restoration, image editing, and virtual reality. Despite the significant advances in face generative models, ensuring that a person's unique facial identity is maintained during the inpainting process is still an elusive goal. Current state-of-the-art techniques, exemplified by MyStyle, necessitate resource-intensive fine-tuning and a substantial number of images for each new identity. Furthermore, existing methods often fall short in accommodating user-specified semantic attributes, such as beard or expression. To improve inpainting results, and reduce the computational complexity during inference, this paper proposes the use of Parallel Visual Attention (PVA) in conjunction with diffusion models. Specifically, we insert parallel attention matrices to each cross-attention module in the denoising network, which attends to features extracted from reference images by an identity encoder. We train the added attention modules and identity encoder on CelebAHQ-IDI, a dataset proposed for identity-preserving face inpainting. Experiments demonstrate that PVA attains unparalleled identity resemblance in both face inpainting and face inpainting with language guidance tasks, in comparison to various benchmarks, including MyStyle, Paint by Example, and Custom Diffusion. Our findings reveal that PVA ensures good identity preservation while offering effective language-controllability. Additionally, in contrast to Custom Diffusion, PVA requires just 40 fine-tuning steps for each new identity, which translates to a significant speed increase of over 20 times.

****

AvatarOne: Monocular 3D Human Animation

Akash Karthikeyan, Robert Ren, Yash Kant, Igor Gilitschenski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3647-3657

Reconstructing realistic human avatars from monocular videos is a challenge that demands intricate modeling of 3D surface and articulation. In this paper, we introduce a comprehensive approach that synergizes three pivotal components: (1) a Signed Distance Field (SDF) representation with volume rendering and grid-based ray sampling to prune empty raysets, enabling efficient 3D reconstruction; (2) faster 3D surface reconstruction through a warmup stage for human surfaces, which ensures detailed modeling of body limbs; and (3) temporally consistent subject specific forward canonical skinning, which helps in retaining correspondences across frames, all of which can be trained in an end-to-end fashion under 30 mins.

Leveraging warmup and grid-based ray marching, along with a faster voxel-based correspondence search, our model streamlines the computational demands of the problem. We further experiment with different sampling representations to improve ray radiance approximations and obtain a floater free surface. Through rigorous evaluation, we demonstrate that our method is on par with current techniques while offering novel insights and avenues for future research in 3D avatar modeling. This work showcases a fast and robust solution for both surface modeling and novel view animation.

****

Synthesizing Anyone, Anywhere, in Any Pose
Håkon Hukkelås, Frank Lindseth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4035-4046

We address the task of in-the-wild human figure synthesis, where the primary goal is to synthesize a full body given any region in any image. In-the-wild human figure synthesis has long been a challenging and under-explored task, where current methods struggle to handle extreme poses, occluding objects, and complex backgrounds. Our main contribution is TriA-GAN, a keypoint-guided GAN that can synthesize Anyone, Anywhere, in Any given pose. Key to our method is projected GANs combined with a well-crafted training strategy, where our simple generator architecture can successfully handle the challenges of in-the-wild full-body synthesis. We show that TriA-GAN significantly improves over previous in-the-wild full-body synthesis methods, all while requiring less conditional information for synthesis (keypoints v.s. DensePose). Finally, we show that the latent space of TriA-GAN is compatible with standard unconditional editing techniques, enabling text-guided editing of generated human figures.

****

Ray Deformation Networks for Novel View Synthesis of Refractive Objects
Weijian Deng, Dylan Campbell, Chunyi Sun, Shubham Kanitkar, Matthew Shaffer, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3118-3128

Neural Radiance Fields (NeRF) have demonstrated exceptional capabilities in creating photorealistic novel views using volume rendering on a radiance field. However, the intrinsic assumption of straight light rays within NeRF becomes a limitation when dealing with transparent or translucent objects that exhibit refraction, and therefore have curved light paths. This hampers the ability of these approaches to accurately model the appearance of refractive objects, resulting in suboptimal novel view synthesis and geometry estimates. To address this issue, we propose an innovative solution using deformable networks to learn a tailored deformation field for refractive objects. Our approach predicts position and direction offsets, allowing NeRF to model the curved light paths caused by refraction and therefore the complex and highly view-dependent appearances of refractive objects. We also introduce a regularization strategy that encourages piece-wise linear light paths, since most physical systems can be approximated with a piece-wise constant index of refraction. By seamlessly integrating our deformation networks into the NeRF framework, our method achieves significant improvements in rendering refractive objects from novel views.

****

NITEC: Versatile Hand-Annotated Eye Contact Dataset for Ego-Vision Interaction
Thorsten Hempel, Magnus Jung, Ahmed A. Abdelrahman, Ayoub Al-Hamadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4437-4446

Eye contact is a crucial non-verbal interaction modality and plays an important role in our everyday social life. While humans are very sensitive to eye contact, the capabilities of machines to capture a person's gaze are still mediocre. We tackle this challenge and present NITEC, a hand-annotated eye contact dataset for ego-vision interaction. NITEC exceeds existing datasets for ego-vision eye contact in size and variety of demographics, social contexts, and lighting conditions, making it a valuable resource for advancing ego-vision-based eye contact research. Our extensive evaluations on NITEC demonstrate strong cross-dataset performance, emphasizing its effectiveness and adaptability in various scenarios, th at allows seamless utilization to the fields of computer vision, human-computer interaction, and social robotics. We make our NITEC dataset publicly available to foster reproducibility and further exploration in the field of ego-vision interaction.

****

Tunable Hybrid Proposal Networks for the Open World
Matthew Inkawhich, Nathan Inkawhich, Hai Li, Yiran Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1988-1999

Current state-of-the-art object proposal networks are trained with a closed-world assumption, meaning they learn to only detect objects of the training classes. These models fail to provide high recall in open-world environments where important novel objects may be encountered. While a handful of recent works attempt to tackle this problem, they fail to consider that the optimal behavior of a proposal network can vary significantly depending on the data and application. Our goal is to provide a flexible proposal solution that can be easily tuned to suit a variety of open-world settings. To this end, we design a Tunable Hybrid Proposal Network (THPN) that leverages an adjustable hybrid architecture, a novel self-training procedure, and dynamic loss components to optimize the tradeoff between known and unknown object detection performance. To thoroughly evaluate our method, we devise several new challenges which invoke varying degrees of label bias by altering known class diversity and label count. We find that in every task, THPN easily outperforms existing baselines (e.g., RPN, OLN). Our method is also highly data efficient, surpassing baseline recall with a fraction of the labeled data.

****

3D Reconstruction of Interacting Multi-Person in Clothing From a Single Image
Junuk Cha, Hansol Lee, Jaewon Kim, Nhat Nguyen Bao Truong, Jaeshin Yoon, Seungryul Baek; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5303-5312

This paper introduces a novel pipeline to reconstruct the geometry of interacting multi-person in clothing on a globally coherent scene space from a single image. The main challenge arises from the occlusion: a part of a human body is not visible from a single view due to the occlusion by others or the self, which introduces missing geometry and physical implausibility (e.g., penetration). We overcome this challenge by utilizing two human priors for complete 3D geometry and surface contacts. For the geometry prior, an encoder learns to regress the image of a person with missing body parts to the latent vectors; a decoder decodes these vectors to produce 3D features of the associated geometry; and an implicit network combines these features with a surface normal map to reconstruct a complete and detailed 3D humans. For the contact prior, we develop an image-space contact detector that outputs a probability distribution of surface contacts between people in 3D. We use these priors to globally refine the body poses, enabling the penetration-free and accurate reconstruction of interacting multi-person in clothing on the scene space. The results demonstrate that our method is complete, globally coherent, and physically plausible compared to existing methods.

****

LensNeRF: Rethinking Volume Rendering Based on Thin-Lens Camera Model
Min-Jung Kim, Gyojung Gu, Jaegul Choo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3182-3191

Recent advances in Neural Radiance Field (NeRF) show promising results in rendering realistic novel view images. However, NeRF and its variants assume that input images are captured using a pinhole camera and that subjects in images are always all-in-focus by tacit agreement. In this paper, we propose aperture-aware NeRF optimization and rendering methods using a thin-lens model (dubbed LensNeRF), which allows defocus images of any aperture size as input and output. To generalize a pinhole camera model to a thin-lens camera model in NeRF framework, we define multiple rays originating from the aperture area, solving world-to-pixel scale ambiguity. Also, we propose in-focus loss that assigns the given pixel color to points on the focus plane to alleviate the color ambiguity caused by the use of multiple rays. For the rigorous evaluation of the proposed method, we collect a real forward-facing dataset with different F-numbers for each viewpoint. Experimental results demonstrate that our method successfully fuses an aperture-size adjustable thin-lens camera model into the NeRF architecture, showing favorable qualitative and quantitative results compared to baseline models. The dataset will be made available.

****

Composite Diffusion: whole >= Sparts

Vikram Jamwal, Ramaneswaran S.; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7231-7240

For artists or graphic designers, the spatial arrangement of a scene is a critical design choice. However, existing text-to-image diffusion models provide limited support for incorporating spatial information. This paper introduces Composite Diffusion as a means for artists to generate high-quality images by composing from sub-scenes. The artists can specify the arrangement of the sub-scenes through a free-form segment layout and can describe the content of each sub-scene using natural text and additional control inputs. We provide a comprehensive and modular framework for Composite Diffusion that enables alternative ways of generating, composing, and harmonizing sub-scenes. We further argue that existing image quality metrics lack a holistic evaluation of image composites. To address this, we propose novel quality criteria especially relevant to composite generation. We believe that our approach provides an intuitive method of art creation. Through extensive user surveys and quantitative and qualitative analysis, we show how it achieves greater spatial, semantic, and creative control over image generation. In addition, our methods do not need to retrain or modify the architecture of the base diffusion models and can work in a plug-and-play manner with the fine-tuned models.

****

P2D: Plug and Play Discriminator for Accelerating GAN Frameworks

Min Jin Chong, Krishna Kumar Singh, Yijun Li, Jingwan Lu, David Forsyth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5422-5431

Most image classification tasks benefit from using pretrained feature stacks. In contrast, the discriminator for adversarial losses is trained at the same time as the model because using a pretrained feature stack yields a very poor model. Recent work has shown that an implicit regularization scheme allows using pretrained feature stacks to construct a discriminator, which improves both speed of training and quality of results. However, we observe that changes in hyperparameters can result in substantial changes in generator behavior. We show that using a modified version of the R1 regularization scheme that regularizes in the feature space instead of the image space results in a plug-and-play discriminator -- P2D. Our scheme results in a method that is highly stable across changes in architecture and framework; that significantly speeds up training; and that produces models that reliably beat SOTA in quality. The huge reduction in training resources required means that P2D could make training powerful generative models over specific datasets accessible to most researchers.

****

PMI Sampler: Patch Similarity Guided Frame Selection for Aerial Action Recognition

Ruiqi Xian, Xijun Wang, Divya Kothandaraman, Dinesh Manocha; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6982-6991

We present a new algorithm for the selection of informative frames in video action recognition. Our approach is designed for aerial videos captured using a moving camera where human actors occupy a small spatial resolution of video frames. Our algorithm utilizes the motion bias within aerial videos, which enables the selection of motion-salient frames. We introduce the concept of patch mutual information (PMI) score to quantify the motion bias between adjacent frames, by measuring the similarity of patches. We use this score to assess the amount of discriminative motion information contained in one frame relative to another. We pres ent an adaptive frame selection strategy using shifted leaky ReLu and cumulative distribution function, which ensures that the sampled frames comprehensively cover all the essential segments with high motion salience. Our approach can be integrated with any action recognition model to enhance its accuracy. In practice, our method achieves a relative improvement of 2.2 - 13.8% in top-1 accuracy on UAV-Human, 6.8% on NEC Drone, and 9.0% on Diving48 datasets. The code is available at https://github.com/Ricky- Xian/PMI-Sampler.

****

REALM: Robust Entropy Adaptive Loss Minimization for Improved Single-Sample Test-Time Adaptation

Skyler Seto, Barry-John Theobald, Federico Danieli, Navdeep Jaitly, Dan Busbridge; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2062-2071

Fully-test-time adaptation (F-TTA) can mitigate performance loss due to distribution shifts between train and test data (1) without access to the training data, and (2) without knowledge of the model training procedure. In online F-TTA, a pre-trained model is adapted using a stream of test samples by minimizing a self-supervised objective, such as entropy minimization. However, models adapted with online using entropy minimization, are unstable especially in single sample settings, leading to degenerate solutions, and limiting the adoption of TTA inference strategies. Prior works identify noisy, or unreliable, samples as a cause of failure in online F-TTA. One solution is to ignore these samples, which can lead to bias in the update procedure, slow adaptation, and poor generalization. In this work, we present a general framework for improving robustness of F-TTA to these noisy samples, inspired by self-paced learning and robust loss functions. Our proposed approach, Robust Entropy Adaptive Loss Minimization (REALM), achieves better adaptation accuracy than previous approaches throughout the adaptation process on corruptions of CIFAR-10 and ImageNet-1K, demonstrating its effectiveness.

****

TSA2: Temporal Segment Adaptation and Aggregation for Video Harmonization

Zeyu Xiao, Yurui Zhu, Xueyang Fu, Zhiwei Xiong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4136-4145

Video composition merges the foreground and background of different videos, presenting challenges due to variations in capture conditions (e.g., saturation, brightness, and contrast). Video harmonization is a vital process in achieving a realistic composite by seamlessly adjusting the foreground's appearance to match the background. In this paper, we propose TSA2, a novel method for video harmonization that incorporates temporal segment adaptation and aggregation. TSA2 divides the inharmonious input sequence into temporal segments, each corresponding to a different frame rate, allowing effective utilization of complementary information within each segment. The method includes the Temporal Segment Adaptation module, which learns and remaps the distribution difference between background and foreground regions, and the Temporal Segment Aggregation module, which emphasizes and aggregates cross-segment information through element-wise correlations. Experimental results demonstrate that TSA2 outperforms advanced image and video harmonization methods quantitatively and qualitatively.

****

PMVC: Promoting Multi-View Consistency for 3D Scene Reconstruction

Chushan Zhang, Jinguang Tong, Tao Jun Lin, Chuong Nguyen, Hongdong Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3678-3688

Reconstructing the geometry of a 3D scene from its multi-view 2D observations has been a central task of 3D computer vision. Recent methods based on neural rendering that use implicit shape representations, such as the neural Signed Distance Function(SDF), have shown impressive performance. However, they fall short in recovering fine details in the scene, especially when employing an MLP as the interpolation function for the SDF representation. Per-frame image normal or depth-map prediction have been utilized to tackle this issue, but these learning-based depth/normal predictions are based on a single image frame only, hence overloo king the underlying multiview consistency of the scene, leading to inconsistent erroneous 3D reconstruction. To mitigate this problem, we propose to leverage multi-view deep features computed on the images. In addition, we employ an adaptive sampling strategy that assesses the fidelity of the multi-view image consistency. Our approach outperforms current state-of-the-art methods, delivering an accurate and robust scene representation with particularly enhanced details in those thin or textureless regions. The effectiveness of our proposed approach is evaluated by extensive experiments conducted on the ScanNet and Replica datasets, showing superior performance than the current state-of-the-art.

****

MGM-AE: Self-Supervised Learning on 3D Shape Using Mesh Graph Masked Autoencoders

Zhangsihao Yang, Kaize Ding, Huan Liu, Yalin Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3303-3313

The challenges of applying self-supervised learning to 3D mesh data include difficulties in explicitly modeling and leveraging geometric topology information and designing appropriate pretext tasks and augmentation methods for irregular mesh topology. In this paper, we propose a novel approach for pre-training models on large-scale, unlabeled datasets using graph masking on a mesh graph composed of faces. Our method, Mesh Graph Masked Autoencoders (MGM-AE), utilizes masked autoencoding to pre-train the model and extract important features from the data. Our pre-trained model outperforms prior state-of-the-art mesh encoders in shape classification and segmentation benchmarks, achieving 90.8% accuracy on ModelNet40 and 78.5 mIoU on ShapeNet. The best performance is obtained when the model is trained and evaluated under different masking ratios. Our approach demonstrates effectiveness in pre-training models on large-scale, unlabeled datasets and its potential for improving performance on downstream tasks.

****

Interactive Network Perturbation Between Teacher and Students for Semi-Supervised Semantic Segmentation

Hyuna Cho, Injun Choi, Suha Kwak, Won Hwa Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 626-635

The current golden standard of semi-supervised semantic segmentation is to generate and exploit pseudo-supervision on unlabeled images. This approach is however susceptible to the quality of pseudo-supervision--training often becomes unstable particularly at early stages and biased to incorrect supervision. To address these issues, we propose a new semi-supervised learning framework, dubbed Guided Pseudo Supervision (GPS). GPS comprises three networks, i.e., a teacher and two separate students. The teacher is first trained with a small set of labeled data and provides stable initial pseudo-supervision on the unlabeled data to the students. The students interactively train each other under the supervision of the teacher, and once they are sufficiently trained, they offer feedback supervision to the teacher so that the teacher improves in subsequent iterations. This strategy enables more stable and faster convergence than previous works, and consequently, GPS achieved state-of-the-art performance on Pascal VOC 2012 and Cityscapes datasets in various experiment settings.

****

Minimizing Layerwise Activation Norm Improves Generalization in Federated Learning

M. Yashwanth, Gaurav Kumar Nayak, Harsh Rangwani, Arya Singh, R. Venkatesh Babu, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2287-2296

Federated Learning (FL) is an emerging machine learning framework that enables multiple clients (coordinated by a server) to collaboratively train a global model by aggregating the locally trained models without sharing any client's training data. It has been observed in recent works that learning in a federated manner may lead the aggregated global model to converge to a 'sharp minimum' thereby adversely affecting the generalizability of this FL-trained model. Therefore, in this work, we aim to improve the generalization performance of models trained in a federated setup by introducing a 'flatness' constrained FL optimization probl em. This flatness constraint is imposed on the top eigenvalue of the Hessian computed from the training loss. %of the global model. As each client trains a model on its local data, we further re-formulate this complex problem utilizing the client loss functions and propose a new computationally efficient regularization technique, dubbed 'MAN' which Minimizes Activation's Norm of each layer on client-side models. We also theoretically show that minimizing the activation norm reduces the top eigenvalue of the layer-wise Hessian of the client's loss, which in turn decreases the overall Hessian's top eigenvalue, ensuring convergence to a flat minimum. We apply our proposed flatness-constrained optimization to the existing FL techniques and obtain significant improvements, thereby establishing new state-of-the-art.

****

ReCLIP: Refine Contrastive Language Image Pre-Training With Source Free Domain Adaptation

Xuefeng Hu, Ke Zhang, Lu Xia, Albert Chen, Jiajia Luo, Yuyin Sun, Ken Wang, Nan Qiao, Xiao Zeng, Min Sun, Cheng-Hao Kuo, Ram Nevatia; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2994-3003

Large-scale pre-training vision-language models (VLM) such as CLIP has demonstrated outstanding performance in zero-shot classification, e.g. achieving 76.3% top-1 accuracy on ImageNet without seeing any example, which leads to potential benefits to many tasks that have no labeled data. However, while applying CLIP to a downstream target domain, the presence of visual and text domain gaps and cross-modality misalignment can greatly impact the model performance. To address such challenges, we propose ReCLIP, a novel source-free domain adaptation method for vision-language models, which does not require any source data or target labeled data. ReCLIP first learns a projection space to mitigate the misaligned visual-text embeddings and learns pseudo labels, and then deploys cross-modality self-training with the pseudo labels, to update visual and text encoders, refine labels and reduce domain gaps and misalignment iteratively. With extensive experiments, we demonstrate that ReCLIP outperforms all the baselines with significant margin and improves the averaged accuracy of CLIP from 69.83% to 74.94% on 22 image classification benchmarks.

****

PointCT: Point Central Transformer Network for Weakly-Supervised Point Cloud Semantic Segmentation

Anh-Thuan Tran, Hoanh-Su Le, Suk-Hwan Lee, Ki-Ryong Kwon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3556-3565

Although point cloud segmentation has a principal role in 3D understanding, annotating fully large-scale scenes for this task can be costly and time-consuming. To resolve this issue, we propose Point Central Transformer (PointCT), a novel end-to-end trainable transformer network for weakly-supervised point cloud semantic segmentation. Divergent from prior approaches, our method addresses limited point annotation challenges exclusively based on 3D points through central-based attention. By employing two embedding processes, our attention mechanism integrates global features across neighborhoods, thereby effectively enhancing unlabeled point representations. Simultaneously, the interconnections between central points and their distinct neighborhoods are bidirectional cohered. Position encoding is further applied to enforce geometric features and improve overall performance. Notably, PointCT achieves outstanding performance under various labeled point settings without additional supervision. Extensive experiments on public datasets S3DIS, ScanNet-V2, and STPLS3D demonstrate the superiority of our proposed approach over other state-of-the-art methods.

****

Lightweight Delivery Detection on Doorbell Cameras

Pirazh Khorramshahi, Zhe Wu, Tianchen Wang, Luke DeLuccia, Hongcheng Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6962-6971

Despite recent advances in video-based action recognition and robust spatio-temp

oral modeling, most of the proposed approaches rely on the abundance of computational resources to afford running huge and computation-intensive convolutional or transformer-based neural networks to obtain satisfactory results. This limits the deployment of such models on edge devices with limited power and computing resources. In this work we investigate an important smart home application, video based delivery detection, and present a simple and lightweight pipeline for this task that can run on resource-constrained doorbell cameras. Our proposed pipeline relies on motion cues to generate a set of coarse activity proposals followed by their classification with a mobile-friendly 3DCNN network. For training we design a novel semi-supervised attention module that helps the network to learn robust spatio-temporal features and adopt an evidence-based optimization objective that allows for quantifying the uncertainty of predictions made by the network. Experimental results on our curated delivery dataset shows the significant effectiveness of our pipeline compared to alternatives and highlights the benefits of our training phase novelties to achieve free and considerable inference-time performance gains.
****
Designing a Hybrid Neural System To Learn Real-World Crack Segmentation From Fractal-Based Simulation
Achref Jaziri, Martin Mundt, Andres Fernandez, Visvanathan Ramesh; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8636-8646
Identification of cracks is essential to assess the structural integrity of concrete infrastructure. However, robust crack segmentation remains a challenging task for computer vision systems due to the diverse appearance of concrete surfaces, variable lighting and weather conditions, and the overlapping of different defects. In particular recent data-driven methods struggle with the limited availability of data, the fine-grained and time-consuming nature of crack annotation, and face subsequent difficulty in generalizing to out-of-distribution samples. In this work, we move past these challenges in a two-fold way. We introduce a high-fidelity crack graphics simulator based on fractals and a corresponding fully-annotated crack dataset. We then complement the latter with a system that learns generalizable representations from simulation, by leveraging both a pointwise mutual information estimate along with adaptive instance normalization as inductive biases. Finally, we empirically highlight how different design choices are symbiotic in bridging the simulation to real gap, and ultimately demonstrate that our introduced system can effectively handle real-world crack segmentation.
****
Deep Visual-Genetic Biometrics for Taxonomic Classification of Rare Species
Tayfun Karaderi, Tilo Burghardt, Raphaël Morard, Daniela N. Schmidt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7115-7125
Visual as well as genetic biometrics are routinely employed to identify species and individuals in biological applications. However, no attempts have been made in this domain to computationally enhance visual classification of rare classes with little image data via genetics. In this paper, we thus propose aligned visual-genetic learning as a new application domain with the aim to implicitly encode cross-modality associations for improved performance. We demonstrate for the first time that such alignment can be achieved via deep embedding models and that the approach is directly applicable to boosting long-tailed recognition (LTR), particularly for rare species. We experimentally demonstrate the efficacy of the concept via application to microscopic imagery of 30k+ planktic foraminifer shells across 32 species when used together with independent genetic data samples. Most importantly for practitioners, we show that visual-genetic alignment can significantly benefit visual-only recognition of the rarest species. Technically, we pre-train a visual ResNet50 deep learning model using triplet loss formulations to create an initial embedding space. We re-structure this space based on genetic anchors embedded via a Sequence Graph Transform (SGT) and linked to visual data by cross-domain cosine alignment. We show that an LTR approach improves the state-of-the-art across all benchmarks and that adding our visual-genetic align

ment improves per-class and particularly rare tail class benchmarks significantly further. We conclude that visual-genetic alignment can be a highly effective tool for complementing visual biological data containing rare classes. The concept proposed may serve as an important future tool for integrating genetics and imageomics towards a more complete scientific representation of taxonomic spaces and life itself. Code, weights, and data splits are published for full reproducibility.
****
Handformer2T: A Lightweight Regression-Based Model for Interacting Hands Pose Estimation From a Single RGB Image
Pengfei Zhang, Deying Kong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6248-6257
Despite its extensive range of potential applications in virtual reality and augmented reality, 3D hand pose estimation from RGB image remains a very challenging problem. The appearance confusions between the two hands and their joints, along with severe hand-hand occlusion and self-occlusion, makes it even more difficult in the senario of interacting hands. Previous methods deal with this problem at the joint level and generally use a heatmap-based method for coordinate prediction. In this paper, we propose a regression-based method that can deal with joint regression at the hand level, which makes the model much more lightweight and memory efficient. To achieve this, we design a novel Pose Query Enhancer (PQE) module, which takes the coarse joint prediction for each hand and refine the prediction iteratively. The key idea of PQE is to make the regression model focus more on the information near proposed joint prediction by manually sampling the feature map. Since we always adopt the transformer on hand level, our model remains lightweight amd memory friendly with this module. Experiments on public benchmarks demonstrate that our model achieves state-of-the-art performance with higher throughput, while requiring less memory and time.
****
Universal Semi-Supervised Model Adaptation via Collaborative Consistency Training
Zizheng Yan, Yushuang Wu, Yipeng Qin, Xiaoguang Han, Shuguang Cui, Guanbin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 872-882
In this paper, we introduce a realistic and challenging domain adaptation problem called Universal Semi-supervised Model Adaptation (USMA), which i) requires only a pre-trained source model, ii) allows the source and target domain to have different label sets, i.e., they share a common label set and hold their own private label set, and iii) requires only a few labeled samples in each class of the target domain. To address USMA, we propose a collaborative consistency training framework that regularizes the prediction consistency between two models, i.e., a pre-trained source model and its variant pre-trained with target data only, and combines their complementary strengths to learn a more powerful model. The rationale of our framework stems from the observation that the source model performs better on common categories than the target-only model, while on target-private categories, the target-only model performs better. We also propose a two-perspective, i.e., sample-wise and class-wise, consistency regularization to improve the training. Experimental results demonstrate the effectiveness of our method on several benchmark datasets.
****
Universal Test-Time Adaptation Through Weight Ensembling, Diversity Weighting, and Prior Correction
Robert A. Marsden, Mario Döbler, Bin Yang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2555-2565
Since distribution shifts are likely to occur during test-time and can drastically decrease the model's performance, online test-time adaptation (TTA) continues to update the model after deployment, leveraging the current test data. Clearly, a method proposed for online TTA has to perform well for all kinds of environmental conditions. By introducing the variable factors domain non-stationarity and temporal correlation, we first unfold all practically relevant settings and de

fine the entity as universal TTA. We want to highlight that this is the first work that covers such a broad spectrum, which is indispensable for the use in practice. To tackle the problem of universal TTA, we identify and highlight several challenges a self-training based method has to deal with: 1) model bias and the occurrence of trivial solutions when performing entropy minimization on varying sequence lengths with and without multiple domain shifts, 2) loss of generalization which exacerbates the adaptation to multiple domain shifts and the occurrence of catastrophic forgetting, and 3) performance degradation due to shifts in class prior. To prevent the model from becoming biased, we leverage a dataset and model-agnostic certainty and diversity weighting. In order to maintain generalization and prevent catastrophic forgetting, we propose to continually weight-average the source and adapted model. To compensate for disparities in the class prior during test-time, we propose an adaptive prior correction scheme that reweights the model's predictions. We evaluate our approach, named ROID, on a wide range of settings, datasets, and models, setting new standards in the field of universal TTA. Code is available at: https://github.com/mariodoebler/test-time-adaptation

****

Uncertainty Estimation in Instance Segmentation With Star-Convex Shapes

Qasim M. K. Siddiqui, Sebastian Starke, Peter Steinbach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1424-1433

Instance segmentation has witnessed promising advancements through deep neural network-based algorithms. However, these models often exhibit incorrect predictions with unwarranted confidence levels. Consequently, evaluating prediction uncertainty becomes critical for informed decision-making. Existing methods primarily focus on quantifying uncertainty in classification or regression tasks, lacking emphasis on instance segmentation. Our research addresses the challenge of estimating spatial certainty associated with the location of instances with star-convex shapes. Two distinct clustering approaches are evaluated which compute spatial and fractional certainty per instance employing samples by the Monte-Carlo Dropout or Deep Ensemble technique. Our study demonstrates that combining spatial and fractional certainty scores yields improved calibrated estimation over individual certainty scores. Notably, our experimental results show that the Deep Ensemble technique alongside our novel radial clustering approach proves to be an effective strategy. Our findings emphasize the significance of evaluating the calibration of estimated certainties for model reliability and decision-making.

****

Spatio-Temporal Filter Analysis Improves 3D-CNN for Action Classification

Takumi Kobayashi, Jiaxing Ye; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6972-6981

As 2D-CNNs are growing in image recognition literature, 3D-CNNs are enthusiastically applied to video action recognition. While spatio-temporal (3D) convolution successfully stems from spatial (2D) convolution, it is still unclear how the convolution works for encoding temporal motion patterns in 3D-CNNs. In this paper, we shed light on the mechanism of feature extraction through analyzing the spatio-temporal filters from a temporal viewpoint. The analysis not only describes characteristics of the two action datasets, Something-Something-v2 (SSv2) and Kinetics-400, but also reveals how temporal dynamics are characterized through stacked spatio-temporal convolutions. Based on the analysis, we propose methods to improve temporal feature extraction, covering temporal filter representation and temporal data augmentation. The proposed method contributes to enlarging temporal receptive field of 3D-CNN without touching its fundamental architecture, thus keeping the computation cost. In the experiments on action classification using SSv2 and Kinetics-400, it produces favorable performance improvement of 3D-CNNs.

****

Bipartite Graph Diffusion Model for Human Interaction Generation

Baptiste Chopin, Hao Tang, Mohamed Daoudi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5333-5342

The generation of natural human motion interactions is a hot topic in computer vision and computer animation. It is a challenging task due to the diversity of possible human motion interactions. Diffusion models, which have already shown remarkable generative capabilities in other domains, are a good candidate for this task. In this paper, we introduce a novel bipartite graph diffusion method (BiGraphDiff) to generate human motion interactions between two persons. Specifically, bipartite node sets are constructed to model the inherent geometric constraints between skeleton nodes during interactions. The interaction graph diffusion model is transformer-based, combining some state-of-the-art motion methods. We show that the proposed achieves new state-of-the-art results on leading benchmarks for the human interaction generation task.

****

Latent Feature-Guided Diffusion Models for Shadow Removal

Kangfu Mei, Luis Figueroa, Zhe Lin, Zhihong Ding, Scott Cohen, Vishal M. Patel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4313-4322

Recovering textures beneath shadows has remained a challenging problem due to the inherent difficulty of inferring shadow-free scenes from shadow images. In this paper, we propose the use of diffusion models as they offer a promising approach to gradually refine details of shadow regions during the diffusion process. Our method improves the process by conditioning on a learned latent feature space that inherits the characteristics of shadow-free images, which has been a limitation of conventional methods that condition on degraded images only. Additionally, we propose to alleviate the potential local optimum of model optimization by fusing noise features with the diffusion network. We demonstrate the effectiveness of our approach, where it outperforms the previous best method by 13% in terms of RMSE on the AISTD dataset and outperforms the previous best method by 82% in terms of RMSE on the DeSOBA dataset for instance-level shadow removal.

****

HAMMER: Learning Entropy Maps To Create Accurate 3D Models in Multi-View Stereo

Rafael Weilharter, Friedrich Fraundorfer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3466-3475

While the majority of recent Multi-View Stereo Networks estimates a depth map per reference image, their performance is then only evaluated on the fused 3D model obtained from all images. This approach makes a lot of sense since ultimately the point cloud is the result we are mostly interested in. On the flip side, it often leads to a burdensome manual search for the right fusion parameters in order to score well on the public benchmarks. In this work, we tackle the aforementioned problem with HAMMER, a Hierarchical And Memory-efficient MVSNet with Entropy-filtered Reconstructions. We propose to learn a filtering mask based on entropy, which, in combination with a simple two-view geometric verification, is sufficient to generate high quality 3D models of any input scene. Distinct from existing works, a tedious manual parameter search for the fusion step is not required. Furthermore, we take several precautions to keep the memory requirements for our method very low in the training as well as in the inference phase. Our method only requires 6 GB of GPU memory during training, while 3.6 GB are enough to process 1920 x 1024 images during inference. Experiments show that HAMMER ranks amongst the top published methods on the DTU and Tanks and Temples benchmarks in the official metrics, especially when keeping the fusion parameters fixed.

****

Localization and Manipulation of Immoral Visual Cues for Safe Text-to-Image Generation

Seongbeom Park, Suhong Moon, Seunghyun Park, Jinkyu Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4675-4684

Current text-to-image generation methods produce high-resolution and high-quality images, but they should not produce immoral images that may contain inappropriate content from the perspective of commonsense morality. Conventional approaches, however, often neglect these ethical concerns, and existing solutions are often limited to ensure moral compatibility. To address this, we propose a novel me thod that has three main capabilities: (1) our model recognizes the degree of visual commonsense immorality of a given generated image, (2) our model localizes immoral visual (and textual) attributes that make the image visually immoral, and (3) our model manipulates such immoral visual cues into a morally-qualifying alternative. We conduct experiments with various text-to-image generation models, including the state-of-the-art Stable Diffusion model, demonstrating the efficacy of our ethical image manipulation approach. Our human study further confirms that ours is indeed able to generate morally-satisfying images from immoral ones.

****

Dynamic Token-Pass Transformers for Semantic Segmentation

Yuang Liu, Qiang Zhou, Jing Wang, Zhibin Wang, Fan Wang, Jun Wang, Wei Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1827-1836

Vision transformers (ViT) usually extract features via forwarding all the tokens in the self-attention layers from top to toe. In this paper, we introduce dynamic token-pass vision transformers (DoViT) for semantic segmentation, which can adaptively reduce the inference cost for images with different complexity. DoViT gradually stops partial easy tokens from self-attention calculation and keeps the hard tokens forwarding until meeting the stopping criteria. We employ lightweight auxiliary heads to make the token-pass decision and divide the tokens into keeping/stopping parts. With a token separate calculation, the self-attention layers are speeded up with sparse tokens and still work friendly with hardware. A token reconstruction module is built to collect and reset the grouped tokens to their original position in the sequence, which is necessary to predict correct semantic masks. We conduct extensive experiments on two common semantic segmentation tasks, and demonstrate that our method greatly reduces about 40%   60% FLOPs and the drop of mIoU is within 0.8% for various segmentation transformers. The throughput and inference speed of ViT-L/B are increased to more than 2x on Cityscapes.

****

MIST: Medical Image Segmentation Transformer With Convolutional Attention Mixing (CAM) Decoder

Md Motiur Rahman, Shiva Shokouhmand, Smriti Bhatt, Miad Faezipour; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 404-413

One of the common and promising deep learning approaches used for medical image segmentation is transformers, as they can capture long-range dependencies among the pixels by utilizing self-attention. Despite being successful in medical image segmentation, transformers face limitations in capturing local contexts of pixels in multimodal dimensions. We propose a Medical Image Segmentation Transformer (MIST) incorporating a novel Convolutional Attention Mixing (CAM) decoder to address this issue. MIST has two parts- a pre-trained multi-axis vision transformer (MaxViT) is used as an encoder, and the encoded feature representation is passed through the CAM decoder for segmenting the images. In the CAM decoder, an attention-mixer combining multi-head self-attention, spatial attention, and squeeze and excitation attention modules is introduced to capture long-range dependencies in all spatial dimensions. Moreover, to enhance spatial information gain, deep and shallow convolutions are used for feature extraction and receptive field expansion, respectively. The integration of low-level and high-level features from different network stages is enabled by skip connection, allowing MIST to suppress unnecessary information. The experiments show that our MIST transformer with CAM decoder outperforms the state-of-the-art models specifically designed for medical image segmentation on the ACDC and Synapse datasets. Our results also demonstrate that adding the CAM decoder with a hierarchical transformer improves the segmentation performance significantly. Our model with data and code is publicly available on GitHub.

****

Active Learning for Single-Stage Object Detection in UAV Images

Asma Yamani, Albandari Alyami, Hamzah Luqman, Bernard Ghanem, Silvio Giancola; P roceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1860-1869

Unmanned aerial vehicles (UAVs) are widely used for image acquisition in various applications, and object detection is a crucial task for UAV imagery analysis. However, training accurate object detectors requires a large amount of annotated data, which can be expensive and time-consuming. To address this issue, we propose an active learning framework for single-stage object detectors in UAV images. First, we introduce Diverse Uncertainty Aggregation (DUA), a novel uncertainty aggregation method that aims to select images with a more diverse variety of object classes with high uncertainties. Second, we address the problem of class imbalance by adjusting the uncertainty calculation based on the performance of each class. Third, we illustrate how reducing the number of images for labeling does not necessarily lead to a lower labeling cost. Evaluation of our approach on a common UAV dataset shows that we can perform similarly (within 0.02 0.5mAP) to using the whole dataset while using only 25% of the images and 32% of the labeled objects. It also outperforms Random Selection and some other aggregation methods. Evaluation on VOC2012 show also consistent results utilizing only 25% of the labeling cost to reach a performance within 0.1 0.5mAP of using the whole dataset. Our results suggest that our proposed active learning framework can effectively reduce the annotation cost while improving the performance of single-stage object detectors in UAV image settings. The code is available on: https://github.com/asmayamani/DUA

****

WaveMixSR: Resource-Efficient Neural Network for Image Super-Resolution

Pranav Jeevan, Akella Srinidhi, Pasunuri Prathiba, Amit Sethi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5884-5892

Image super-resolution research recently has been dominated by transformer models which need higher computational resources than CNNs due to the quadratic complexity of self-attention. We propose a new neural network -- WaveMixSR -- for image super-resolution based on the WaveMix architecture which uses a 2D-discrete wavelet transform for spatial token-mixing. Unlike transformer-based models, WaveMixSR does not unroll the image as a sequence of pixels/patches. It uses the inductive bias of convolutions along with the lossless token-mixing property of wavelet transform to achieve higher performance while requiring fewer resources and training data. We compare the performance of our network with other state-of-the-art methods for image super-resolution. Our experiments show that WaveMixSR achieves competitive performance in all datasets and reaches state-of-the-art performance in the BSD100 dataset on multiple super-resolution tasks. Our model is able to achieve this performance using less training data and computational resources while maintaining high parameter efficiency compared to current state-of-the-art models.

****

Disentangled Pre-Training for Image Matting

Yanda Li, Zilong Huang, Gang Yu, Ling Chen, Yunchao Wei, Jianbo Jiao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 169-178

Image matting requires high-quality pixel-level human annotations to support the training of a deep model in recent literature. Whereas such annotation is costly and hard to scale, significantly holding back the development of the research. In this work, we make the first attempt towards addressing this problem, by proposing a self-supervised pre-training approach that can leverage infinite numbers of data to boost the matting performance. The pre-training task is designed in a similar manner as image matting, where random trimap and alpha matte are generated to achieve an image disentanglement objective. The pre-trained model is then used as an initialisation of the downstream matting task for fine-tuning. Extensive experimental evaluations show that the proposed approach outperforms both the state-of-the-art matting methods and other alternative self-supervised initialisation approaches by a large margin. We also show the robustness of the proposed approach over different backbone architectures. Our project page is availab

le at https://crystraldo.github.io/dpt_mat/.
****

PromptAD: Zero-Shot Anomaly Detection Using Text Prompts

Yiting Li, Adam Goodge, Fayao Liu, Chuan-Sheng Foo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1093-1102

We target the problem of zero-shot anomaly detection, in which a model is pre-trained on a set of seen classes and expected to detect anomalies in other unseen classes at test time. Although providing exceptional results for many anomaly detection (AD) tasks, state-of-the-art AD algorithms catastrophically struggle in zero-shot scenarios. However, if knowledge of additional modalities exist (e.g. text), we can compensate for the lack of visual information and improve the AD performance. In this work, we propose a knowledge-guided learning framework, namely PromptAD, which achieves the compatibility of a abnormality view and a normality view through a dual-branch vision-language decoding network. Concretely, the normality branch establishes a normality profile to exclude anomalies. Meanwhile, the abnormality branch directly models anomaly behaviors provided by natural language. As the two views capture complementary information, we naturally think of the compatibility of them for achieving better performance. Therefore, a cross-view contrastive learning (CCL) s proposed to regularize the intra-view training with additional reference information from the other complementary view, and a cross-view mutual interaction (CMI) strategy further promotes the mutual exploration of useful knowledge from each branch.
****

Random Walks for Temporal Action Segmentation With Timestamp Supervision

Roy Hirsch, Regev Cohen, Tomer Golany, Daniel Freedman, Ehud Rivlin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6614-6624

Temporal action segmentation relates to high-level video understanding, commonly formulated as frame-wise classification of untrimmed videos into predefined actions. Fully-supervised deep-learning approaches require dense video annotations which are time and money consuming. Furthermore, the temporal boundaries between consecutive actions typically are not well-defined, leading to inherent ambiguity and inter-rater disagreement. A promising approach to remedy these limitations is timestamp supervision, requiring only one labeled frame per action instance in a training video. In this work, we reformulate the task of temporal segmentation as a graph segmentation problem with weakly-labeled vertices. We introduce an efficient segmentation method based on random walks on graphs, obtained by solving a sparse system of linear equations. Furthermore, the proposed technique can be employed in any one or combination of the following forms: (1) as a standalone solution for generating dense pseudo-labels from timestamps; (2) as a training loss; (3) as a smoothing mechanism given intermediate predictions. Extensive experiments with three datasets (50Salads, Breakfast, GTEA) show that our method competes with state-of-the-art, and allows the identification of regions of uncertainty around action boundaries.
****

Masked Collaborative Contrast for Weakly Supervised Semantic Segmentation

Fangwen Wu, Jingxuan He, Yufei Yin, Yanbin Hao, Gang Huang, Lechao Cheng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 862-871

This study introduces an efficacious approach, Masked Collaborative Contrast (MCC), to highlight semantic regions in weakly supervised semantic segmentation. MCC adroitly draws inspiration from masked image modeling and contrastive learning to devise a novel framework that induces keys to contract toward semantic regions. Unlike prevalent techniques that directly eradicate patch regions in the input image when generating masks, we scrutinize the neighborhood relations of patch tokens by exploring masks considering keys on the affinity matrix. Moreover, we generate positive and negative samples in contrastive learning by utilizing the masked local output and contrasting it with the global output. Elaborate experiments on commonly employed datasets evidences that the proposed MCC mechanism effectively aligns global and local perspectives within the image, attaining impr essive performance. The source code is available at https://github.com/fwu11/MCC.

****

Critical Gap Between Generalization Error and Empirical Error in Active Learning

Yusuke Kanebako; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2771-2779

Conventional research papers on Active Learning (AL) have conducted evaluations based on the assumption that a large amount of annotated data is available for evaluating model performance apart from the data selected by AL. This evaluation method is not realistic for the setting where AL learns models with few annotation costs. If a large amount of annotated data is available, it should be used for both evaluation and training, not only for evaluation. Therefore, in a realistic model construction using AL, generalization error in the actual production environment should be estimated by cross-validation only using the data selected by AL. However, the data selected by AL tend to be a biased dataset because the data are selected based on some criteria. Therefore, there is a gap between the actual generalization error and the empirical error when conducting cross-validation on the AL-selected data. In addition, if validation is performed using only the selected dataset by AL, it is possible to fail to realize this fatal gap. In this paper, we show that cross-validation using selected data in conventional AL methods either overestimate or underestimate model performance. As a result, we show a significant difference between generalization error and empirical error from cross-validation.

****

Semi-Supervised Scene Change Detection by Distillation From Feature-Metric Alignment

Seonhoon Lee, Jong-Hwan Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1226-1235

Scene change detection (SCD) is a critical task for various applications, such as visual surveillance, anomaly detection, and mobile robotics. Recently, supervised methods for SCD have been developed for urban and indoor environments where input image pairs are typically unaligned due to differences in camera viewpoints. However, supervised SCD methods require pixel-wise change labels and alignment labels for the target domain, which can be both time-consuming and expensive to collect. To tackle this issue, we design an unsupervised loss with regularization methods based on the feature-metric alignment of input image pairs. The proposed unsupervised loss enables the SCD model to jointly learn the flow and the change maps on the target domain. In addition, we propose a semi-supervised learning method based on a distillation loss for the robustness of the SCD model. The proposed learning method is based on the student-teacher structure and incorporates the unsupervised loss of the unlabeled target data and the supervised loss of the labeled synthetic data. Our method achieves considerable performance improvement on the target domain through the proposed unsupervised and distillation loss, using only 10% of the target training dataset without using any labels of the target data.

****

Point-DynRF: Point-Based Dynamic Radiance Fields From a Monocular Video

Byeongjun Park, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3171-3181

Dynamic radiance fields have emerged as a promising approach for generating novel views from a monocular video. However, previous methods enforce the geometric consistency to dynamic radiance fields only between adjacent input frames, making it difficult to represent the global scene geometry and degenerates at the viewpoint that is spatio-temporally distant from the input camera trajectory. To solve this problem, we introduce point-based dynamic radiance fields (Point-DynRF), a novel framework where the geometric information and the volume rendering process are trained by neural point clouds and dynamic radiance fields, respectively. Specifically, we reconstruct neural point clouds directly from geometric proxies and optimize both radiance fields and the geometric proxies using our proposed losses, allowing them to complement each other. We validate the effectiveness

of our method with experiments on the NVIDIA Dynamic Scenes Dataset and several causally captured monocular video clips.
****

Re-VoxelDet: Rethinking Neck and Head Architectures for High-Performance Voxel-Based 3D Detection

Jae-Keun Lee, Jin-Hee Lee, Joohyun Lee, Soon Kwon, Heechul Jung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7503-7512

Currently, widely employed LiDAR-based 3D object detectors adopt grid-based approaches to efficiently handle sparse point clouds. However, during this process, the down-sampled features inevitably lose spatial information, which can hinder the detectors from accurately predicting the location and size of objects. To address this issue, previous researches proposed sophisticatedly designed neck and head modules to effectively compensate for information loss. Inspired by the core insights of previous studies, we propose a novel voxel-based 3D object detector, named as Re-VoxelDet, which combines three distinct components to achieve both good detection capability and real-time performance. First, in order to learn features from diverse perspectives without additional computational costs during inference, we introduce Multi-view Voxel Backbone (MVBackbone). Second, to effectively compensate for abundant spatial and strong semantic information, we design Hierarchical Voxel-guided Auxiliary Neck (HVANeck), which attentively integrate hierarchically generated voxel-wise features with RPN blocks. Third, we present Rotation-based Group Head (RGHead), a simple yet effective head module that is designed with two groups according to the heading direction and aspect ratio of the objects. Through extensive experiments on the Argoverse2, nuScenes, and Waymo Open Dataset, we demonstrate the effectiveness of our approach. Our results significantly outperform existing state-of-the-art methods. We plan to release our model and code in the near future.
****

Motion Matters: Neural Motion Transfer for Better Camera Physiological Measurement

Akshay Paruchuri, Xin Liu, Yulu Pan, Shwetak Patel, Daniel McDuff, Soumyadip Sengupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5933-5942

Machine learning models for camera-based physiological measurement can have weak generalization due to a lack of representative training data. Body motion is one of the most significant sources of noise when attempting to recover the subtle cardiac pulse from a video. We explore motion transfer as a form of data augmentation to introduce motion variation while preserving physiological changes of interest. We adapt a neural video synthesis approach to augment videos for the task of remote photoplethysmography (rPPG) and study the effects of motion augmentation with respect to 1) the magnitude and 2) the type of motion. After training on motion-augmented versions of publicly available datasets, the presented inter-dataset results on five benchmark datasets show improvements of up to 79% over existing inter-dataset results using TS-CAN, a neural rPPG estimation method. Additionally, we demonstrate a 47% improvement over existing results on the PURE dataset using various state-of-the-art methods. Our findings illustrate the usefulness of motion transfer as a data augmentation technique for improving the generalization of models for camera-based physiological sensing. We release our code for using motion transfer as a data augmentation technique on three publicly available datasets, UBFC-rPPG, PURE, and SCAMPS, and models pre-trained on motion-augmented data.
****

Towards Diverse and Consistent Typography Generation

Wataru Shimoda, Daichi Haraguchi, Seiichi Uchida, Kota Yamaguchi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7296-7305

In this work, we consider the typography generation task that aims at producing diverse typographic styling for the given graphic document. We formulate typography generation as a fine-grained attribute generation for multiple text elements and build an autoregressive model to generate diverse typography that matches the input design context. We further propose a simple yet effective sampling approach that respects the consistency and distinction principle of typography so that generated examples share consistent typographic styling across text elements. Our empirical study shows that our model successfully generates diverse typographic designs while preserving a consistent typographic structure.

****

IR-FRestormer: Iterative Refinement With Fourier-Based Restormer for Accelerated MRI Reconstruction

Mohammad Zalbagi Darestani, Vishwesh Nath, Wenqi Li, Yufan He, Holger R. Roth, Ziyue Xu, Daguang Xu, Reinhard Heckel, Can Zhao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7655-7664

Accelerated magnetic resonance imaging (MRI) aims to reconstruct high-quality MR images from a set of under-sampled measurements. State-of-the-art methods for this task use deep learning, which offers high reconstruction accuracy and fast runtimes. In this work, we propose a new state-of-the-art reconstruction model for accelerated MRI reconstruction. Our model is the first to combine the power of deep neural networks with iterative refinement for this task. For the neural network component of our method, we utilize a transformer-based architecture as transformers are state-of-the-art in various image reconstruction tasks. However, a major drawback of transformers which has limited their emergence among the state-of-the-art MRI models is that they are often memory inefficient for high-resolution inputs. To address this limitation, we propose a transformer-based model which uses parameter-free Fourier-based attention modules, achieving 2x more memory efficiency. We evaluate our model on the largest publicly available MRI dataset, the fastMRI dataset, and achieve on-par performance with other state-of-the-art methods on the dataset's leaderboard.

****

Deep Plug-and-Play Nighttime Non-Blind Deblurring With Saturated Pixel Handling Schemes

Hung-Yu Shu, Yi-Hsien Lin, Yi-Chang Lu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1538-1546

Due to the setting of shutter speeds, over-exposed blurry images can often be seen in nighttime photography. Although image deblurring is a classic problem in image restoration, state-of-the-art methods often fail in nighttime cases with saturated pixels. The primary reason is that those pixels are out of the sensor range and thus violate the assumption of the linear blur model. To address this issue, we propose a new nighttime non-blind deblurring algorithm with saturated pixel handling schemes, including a pixel stretching mask, an image segment mask, and a saturation awareness mechanism (SAM). Our algorithm achieves superior results by strategically adjusting mask configurations, making our method robust to various saturation levels. We formulate our task into two new optimization problems and introduce a unified framework based on the plug-and-play alternating direction method of multipliers (PnP-ADMM). We also evaluate our approach qualitatively and quantitatively to demonstrate its effectiveness. The results show that the proposed algorithm recovers sharp latent images with finer details and fewer artifacts than other state-of-the-art deblurring methods.

****

One Style Is All You Need To Generate a Video

Sandeep Manandhar, Auguste Genovesio; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5038-5047

In this paper, we propose a style-based conditional video generative model. We introduce a novel temporal generator based on a set of learned sinusoidal bases. Our method learns dynamic representations of various actions that are independent of image content and can be transferred between different actors. Beyond the significant enhancement of video quality compared to prevalent methods, we demonstrate that the disentangled dynamic and content permit their independent manipulation, as well as temporal GAN-inversion to retrieve and transfer a video motion from one content or identity to another without further preprocessing such as landmark points.

****

Wino Vidi Vici: Conquering Numerical Instability of 8-Bit Winograd Convolution for Accurate Inference Acceleration on Edge

Pierpaolo Mori, Lukas Frickenstein, Shambhavi Balamuthu Sampath, Moritz Thoma, Nael Fasfous, Manoj Rohit Vemparala, Alexander Frickenstein, Christian Unger, Walter Stechele, Daniel Mueller-Gritschneder, Claudio Passerone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 53-62

Winograd-based convolution can reduce the total number of operations needed for convolutional neural network (CNN) inference on edge devices. Most edge hardware accelerators use low-precision, 8-bit integer arithmetic units to improve energy efficiency and latency. This makes CNN quantization a critical step before deploying the model on such an edge device. To extract the benefits of fast Winograd-based convolution and efficient integer quantization, the two approaches must be combined. Research has shown that the transform required to execute convolutions in the Winograd domain results in numerical instability and severe accuracy degradation when combined with quantization, making the two techniques incompatible on edge hardware. This paper proposes a novel training scheme to achieve efficient Winograd-accelerated, quantized CNNs. 8-bit quantization is applied to all the intermediate results of the Winograd convolution without sacrificing task-related accuracy. This is achieved by introducing clipping factors in the intermediate quantization stages as well as using the complex numerical system to improve the transform. We achieve 2.8x and 2.1x reduction in MAC operations on ResNet-20-CIFAR-10 and ResNet-18-ImageNet, respectively, with no accuracy degradation.

****

Leveraging Bitstream Metadata for Fast, Accurate, Generalized Compressed Video Quality Enhancement

Max Ehrlich, Jon Barker, Namitha Padmanabhan, Larry Davis, Andrew Tao, Bryan Catanzaro, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1517-1527

Video compression is a central feature of the modern internet powering technologies from social media to video conferencing. While video compression continues to mature, for many compression settings, quality loss is still noticeable. These settings nevertheless have important applications to the efficient transmission of videos over bandwidth constrained or otherwise unstable connections. In this work, we develop a deep learning architecture capable of restoring detail to compressed videos which leverages the underlying structure and motion information embedded in the video bitstream. We show that this improves restoration accuracy compared to prior compression correction methods and is competitive when compared with recent deep-learning-based video compression methods on rate-distortion while achieving higher throughput. Furthermore, we condition our model on quantization data which is readily available in the bitstream. This allows our single model to handle a variety of different compression quality settings which required an ensemble of models in prior work.

****

ECSIC: Epipolar Cross Attention for Stereo Image Compression

Matthias Wödlinger, Jan Kotera, Manuel Keglevic, Jan Xu, Robert Sablatnig; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3436-3445

In this paper, we present ECSIC, a novel learned method for stereo image compression. Our proposed method compresses the left and right images in a joint manner by exploiting the mutual information between the images of the stereo image pair using a novel stereo cross attention (SCA) module and two stereo context modules. The SCA module performs cross-attention restricted to the corresponding epipolar lines of the two images and processes them in parallel. The stereo context modules improve the entropy estimation of the second encoded image by using the first image as a context. We conduct an extensive ablation study demonstrating the effectiveness of the proposed modules and a comprehensive quantitative and qualitative comparison with existing methods. ECSIC achieves state-of-the-art perf ormance in stereo image compression on the two popular stereo image datasets Cityscapes and InStereo2k while allowing for fast encoding and decoding.

****

FacadeNet: Conditional Facade Synthesis via Selective Editing

Yiangos Georgiou, Marios Loizou, Tom Kelly, Melinos Averkiou; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5384-5393

We introduce FacadeNet, a deep learning approach for synthesizing building facade images from diverse viewpoints. Our method employs a conditional GAN, taking a single view of a facade along with the desired viewpoint information and generates an image of the facade from the distinct viewpoint. To precisely modify view-dependent elements like windows and doors while preserving the structure of view-independent components such as walls, we introduce a selective editing module. This module leverages image embeddings extracted from a pretrained vision transformer Our experiments demonstrated state-of-the-art performance on building facade generation, surpassing alternative methods.

****

VEATIC: Video-Based Emotion and Affect Tracking in Context Dataset

Zhihang Ren, Jefferson Ortega, Yifan Wang, Zhimin Chen, Yunhui Guo, Stella X. Yu, David Whitney; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4467-4477

Human affect recognition has been a significant topic in psychophysics and computer vision. However, the currently published datasets have many limitations. For example, most datasets contain frames that contain only information about facial expressions. Due to the limitations of previous datasets, it is very hard to either understand the mechanisms for affect recognition of humans or generalize well on common cases for computer vision models trained on those datasets. In this work, we introduce a brand new large dataset, the Video-based Emotion and Affect Tracking in Context Dataset (VEATIC), that can conquer the limitations of the previous datasets. VEATIC has 124 video clips from Hollywood movies, documentaries, and home videos with continuous valence and arousal ratings of each frame via real-time annotation. Along with the dataset, we propose a new computer vision task to infer the affect of the selected character via both context and character information in each video frame. Additionally, we propose a simple model to benchmark this new computer vision task. We also compare the performance of the pretrained model using our dataset with other similar datasets. Experiments show the competing results of our pretrained model via VEATIC, indicating the generalizability of VEATIC. Our dataset is available at https://veatic.github.io.

****

SimpliMix: A Simplified Manifold Mixup for Few-Shot Point Cloud Classification

Minmin Yang, Weiheng Chai, Jiyang Wang, Senem Velipasalar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3668-3677

Few-shot learning often assumes that base classes are abundant and diverse with plentiful well-labeled samples for each class. This ensures that models can generalize effectively from a small amount of data by leveraging prior knowledge learned from base classes. This assumption holds for 2D few-shot learning since the benchmark datasets are large and diverse. However, 3D point cloud few-shot benchmarks are low in magnitude and diversity. We conduct experiments and show that many existing methods overlook this issue and suffer from overfitting on base classes, which hinders generalization ability and test performance. To alleviate the overfitting issue, we propose a simplified manifold mixup, referred to as the SimpliMix, which mixes hidden representations and forces the models to learn more generalized features. We incorporate SimpliMix into existing prototype-based models, perform experiments on ModelNet40-FS, ModelNet40-C-FS and ScanObjectNN-FS datasets, and improve the models by a significant margin. We further conduct cross-domain few-shot classification experiments and show that networks with SimpliMix learn more generalized and transferable features and achieve better performance. The code is available at https://github.com/LexieYang/SimpliMix

****

ProxEdit: Improving Tuning-Free Real Image Editing With Proximal Guidance

Ligong Han, Song Wen, Qi Chen, Zhixing Zhang, Kunpeng Song, Mengwei Ren, Ruijiang Gao, Anastasis Stathopoulos, Xiaoxiao He, Yuxiao Chen, Di Liu, Qilong Zhangli, Jindong Jiang, Zhaoyang Xia, Akash Srivastava, Dimitris Metaxas; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4291-4301

DDIM inversion has revealed the remarkable potential of real image editing within diffusion-based methods. However, the accuracy of DDIM reconstruction degrades as larger classifier-free guidance (CFG) scales being used for enhanced editing. Null-text inversion (NTI) optimizes null embeddings to align the reconstruction and inversion trajectories with larger CFG scales, enabling real image editing with cross-attention control. Negative-prompt inversion (NPI) further offers a training-free closed-form solution of NTI. However, it may introduce artifacts and is still constrained by DDIM reconstruction quality. To overcome these limitations, we propose proximal guidance and incorporate it to NPI with cross-attention control. We enhance NPI with a regularization term and inversion guidance, which reduces artifacts while capitalizing on its training-free nature. Additionally, we extend the concepts to incorporate mutual self-attention control, enabling geometry and layout alterations in the editing process. Our method provides an efficient and straightforward approach, effectively addressing real image editing tasks with minimal computational overhead.

****

Diverse Imagenet Models Transfer Better

Niv Nayman, Avram Golbert, Asaf Noy, Lihi Zelnik-Manor; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1914-1925

A commonly accepted hypothesis is that models with higher accuracy on Imagenet perform better on other downstream tasks, leading to much research dedicated to optimizing Imagenet accuracy. Recently this hypothesis has been challenged by evidence showing that self-supervised models transfer better than their supervised counterparts, despite their inferior Imagenet accuracy. This calls for identifying the additional factors, on top of Imagenet accuracy, that make models transferable. In this work we show that high diversity of the filters learnt by the model promotes transferability jointly with Imagenet accuracy. Encouraged by the recent transferability results of self-supervised models, we use a simple procedure to combine self-supervised and supervised pretraining and generate models with both high diversity and high accuracy, and as a result high transferability. We experiment with several architectures and multiple downstream tasks, including both single-label and multi-label classification.

****

SOAP: Cross-Sensor Domain Adaptation for 3D Object Detection Using Stationary Object Aggregation Pseudo-Labelling

Chengjie Huang, Vahdat Abdelzad, Sean Sedwards, Krzysztof Czarnecki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3352-3361

We consider the problem of cross-sensor domain adaptation in the context of LiDAR-based 3D object detection and propose Stationary Object Aggregation Pseudo-labelling (SOAP) to generate high quality pseudo-labels for stationary objects. In contrast to the current state-of-the-art in-domain practice of aggregating just a few input scans, SOAP aggregates entire sequences of point clouds at the input level to reduce the sensor domain gap. Then, by means of what we call quasi-stationary training and spatial consistency post-processing, the SOAP model generates accurate pseudo-labels for stationary objects, closing a minimum of 30.3% domain gap compared to few-frame detectors. Our results also show that state-of-the-art domain adaptation approaches can achieve even greater performance in combination with SOAP, in both the unsupervised and semi-supervised settings.

****

Layer-Wise Auto-Weighting for Non-Stationary Test-Time Adaptation

Junyoung Park, Jin Kim, Hyeongjun Kwon, Ilhoon Yoon, Kwanghoon Sohn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 20

24, pp. 1414-1423

Given the inevitability of domain shifts during inference in real-world applications, test-time adaptation (TTA) is essential for model adaptation after deployment. However, the real-world scenario of continuously changing target distributions presents challenges including catastrophic forgetting and error accumulation. Existing TTA methods for non-stationary domain shifts, while effective, incur excessive computational load, making them impractical for on-device settings. In this paper, we introduce a layer-wise auto-weighting algorithm for continual and gradual TTA that autonomously identifies layers for preservation or concentrated adaptation. By leveraging the Fisher Information Matrix (FIM), we first design the learning weight to selectively focus on layers associated with log-likelihood changes while preserving unrelated ones. Then, we further propose an exponential min-max scaler to make certain layers nearly frozen while mitigating outliers. This minimizes forgetting and error accumulation, leading to efficient adaptation to non-stationary target distribution. Experiments on CIFAR-10C, CIFAR-100C, and ImageNet-C show our method outperforms conventional continual and gradual TTA approaches while significantly reducing computational load, highlighting the importance of FIM-based learning weight in adapting to continuously or gradually shifting target domains.

****

Improving Fairness Using Vision-Language Driven Image Augmentation

Moreno D'Incà, Christos Tzelepis, Ioannis Patras, Nicu Sebe; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4695-4704

Fairness is crucial when training a deep-learning discriminative model, especially in the facial domain. Models tend to correlate specific characteristics (such as age and skin color) with unrelated attributes (downstream tasks), resulting in biases which do not correspond to reality. It is common knowledge that these correlations are present in the data and are then transferred to the models during training. This paper proposes a method to mitigate these correlations to improve fairness. To do so, we learn interpretable and meaningful paths lying in the semantic space of a pre-trained diffusion model (DiffAE) -- such paths being supervised by contrastive text dipoles. That is, we learn to edit protected characteristics (age and skin color). These paths are then applied to augment images to improve the fairness of a given dataset. We test the proposed method on CelebA-HQ and UTKFace on several downstream tasks with age and skin color as protected characteristics. As a proxy for fairness, we compute the difference in accuracy with respect to the protected characteristics. Quantitative results show how the augmented images help the model improve the overall accuracy, the aforementioned metric, and the disparity of equal opportunity. Code is available at: https://github.com/Moreno98/Vision-Language-Bias-Control.

****

SupeRVol: Super-Resolution Shape and Reflectance Estimation in Inverse Volume Rendering

Mohammed Brahimi, Bjoern Haefner, Tarun Yenamandra, Bastian Goldluecke, Daniel Cremers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3139-3149

We propose an end-to-end inverse rendering pipeline called SupeRVol that allows us to recover 3D shape and material parameters from a set of color images in a super-resolution manner. To this end, we represent both the bidirectional reflectance distribution function (BRDF) and the signed distance function (SDF) by multi-layer perceptrons. In order to obtain both the surface shape and its reflectance properties, we revert to a differentiable volume renderer with a physically based illumination model that allows us to decouple reflectance and lighting. This physical model takes into account the effect of the camera's point spread function thereby enabling a reconstruction of shape and material in a super-resolution quality. Experimental validation confirms that SupeRVol achieves state of the art performance in terms of inverse rendering quality. It generates reconstructions that are sharper than the individual input images, making this method ideally suited for 3D modeling from low-resolution imagery.

****

Object Aware Contrastive Prior for Interactive Image Segmentation

Praful Mathur, Shashi Kumar Parwani, Mrinmoy Sen, Roopa Sheshadri, Aman Sharma; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 575-584

Interactive Image Segmentation is a process of separating a user selected object from the background. This task requires building an effective class-agnostic segmentation model that performs well even on unseen categories. To achieve good accuracy with limited training dataset, it is important that the model has robust prior understanding of features of similar class objects. The model should also have good distinguishing capabilities of foreground objects with the background. In this paper, we propose Object Aware Click Embeddings (OACE) that represents user click aware foreground object features. OACE is obtained based on a prior network trained using the Contrastive Learning paradigm. The single-click object selection accuracy of our base interactive segmentation network is vastly improved with the OACE input. Additionally, we propose a Multi-Stage fusion approach to better utilize user click information. With the proposed method, we outperform existing state-of-the-art approaches by 21% on publicly available test-sets for click-based Interactive Image Segmentation.

****

Torque Based Structured Pruning for Deep Neural Network

Arshita Gupta, Tien Bau, Joonsoo Kim, Zhe Zhu, Sumit Jha, Hrishikesh Garud; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2711-2720

Structured pruning is a popular way of convolutional neural network (CNN) acceleration. However, current state of the art pruning techniques require modifications to the network architecture, implementation of complex gradient update rules or repetitive training and long fine-tuning stages. Our novel physics-inspired approach for structured pruning aims to solve these issues. Analogous to 'Torque' we apply a force that consolidates the weights of a convolutional layer around a selected pivot point during training. Using the distance-dependency nature of torque, we can encourage high density of weights in filters around this point and increase filter sparsity as we move away. Filters away from the pivot point can be pruned, resulting in a minimum loss of information. We can control the tightness of the weights by varying the hyper-parameters, thus assisting us in creating a more compact network. Our proposed technique is jointly able to perform both filter learning and filter importance sorting. Additionally, our method is easy to implement, requires no change to model architecture and needs very little to no fine-tuning. We show that our approach reaches competitive results with previous state-of-the-art by evaluating popular networks such as VGGNet and ResNet on multiple image classification tasks. Notably, our method can reduce the parameter count of VGGNet by 96% and still maintain the accuracy achieved by the full-size model without any fine-tuning. This makes our method both latency and memory efficient for hardware deployment.

****

Instruct Me More! Random Prompting for Visual In-Context Learning

Jiahao Zhang, Bowen Wang, Liangzhi Li, Yuta Nakashima, Hajime Nagahara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2597-2606

Large-scale models trained on extensive datasets, have emerged as the preferred approach due to their high generalizability across various tasks. In-context learning (ICL), a popular strategy in natural language processing, uses such models for different tasks by providing instructive prompts but without updating model parameters. This idea is now being explored in computer vision, where an input-output image pair (called an in-context pair) is supplied to the model with a query image as a prompt to exemplify the desired output. The efficacy of visual ICL often depends on the quality of the prompts. We thus introduce a method coined Instruct Me More (InMeMo), which augments in-context pairs with a learnable perturbation (prompt), to explore its potential. Our experiments on mainstream tasks reveal that InMeMo surpasses the current state-of-the-art performance. Specifi cally, compared to the baseline without learnable prompt, InMeMo boosts mIoU scores by 7.35 and 15.13 for foreground segmentation and single object detection tasks, respectively. Our findings suggest that InMeMo offers a versatile and efficient way to enhance the performance of visual ICL with lightweight training. Code is available at https://github.com/Jackieam/InMeMo.

****

CL-MAE: Curriculum-Learned Masked Autoencoders

Neelu Madan, Nicolae-C■t■lin Ristea, Kamal Nasrollahi, Thomas B. Moeslund, Radu Tudor Ionescu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2492-2502

Masked image modeling has been demonstrated as a powerful pretext task for generating robust representations that can be effectively generalized across multiple downstream tasks. Typically, this approach involves randomly masking patches (tokens) in input images, with the masking strategy remaining unchanged during training. In this paper, we propose a curriculum learning approach that updates the masking strategy to continually increase the complexity of the self-supervised reconstruction task. We conjecture that, by gradually increasing the task complexity, the model can learn more sophisticated and transferable representations. To facilitate this, we introduce a novel learnable masking module that possesses the capability to generate masks of different complexities, and integrate the proposed module into masked autoencoders (MAE). Our module is jointly trained with the MAE, while adjusting its behavior during training, transitioning from a partner to the MAE (optimizing the same reconstruction loss) to an adversary (optimizing the opposite loss), while passing through a neutral state. The transition between these behaviors is smooth, being regulated by a factor that is multiplied with the reconstruction loss of the masking module. The resulting training procedure generates an easy-to-hard curriculum. We train our Curriculum-Learned Masked Autoencoder (CL-MAE) on ImageNet and show that it exhibits superior representation learning capabilities compared to MAE. The empirical results on five downstream tasks confirm our conjecture, demonstrating that curriculum learning can be successfully used to self-supervise masked autoencoders. We release our code at https://github.com/ristea/cl-mae.

****

Think Before You Simulate: Symbolic Reasoning To Orchestrate Neural Computation for Counterfactual Question Answering

Adam Ishay, Zhun Yang, Joohyung Lee, Ilgu Kang, Dongjae Lim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6698-6707

Causal and temporal reasoning about video dynamics is a challenging problem. While neuro-symbolic models that combine symbolic reasoning with neural-based perception and prediction have shown promise, they exhibit limitations, especially in answering counterfactual questions. This paper introduces a method to enhance a neuro-symbolic model for counterfactual reasoning, leveraging symbolic reasoning about causal relations among events. We define the notion of a causal graph to represent such relations and use Answer Set Programming (ASP), a declarative logic programming method, to find how to coordinate perception and simulation modules. We validate the effectiveness of our approach on two benchmarks, CLEVRER and CRAFT. Our enhancement achieves state-of-the-art performance on the CLEVRER challenge, significantly outperforming existing models. In the case of the CRAFT benchmark, we leverage a large pre-trained language model, such as GPT-3.5 and GPT-4, as a proxy for a dynamics simulator. Our findings show that this method can further improve its performance on counterfactual questions by providing alternative prompts instructed by symbolic causal reasoning.

****

Robust Object Detection in Challenging Weather Conditions

Himanshu Gupta, Oleksandr Kotlyar, Henrik Andreasson, Achim J. Lilienthal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7523-7532

Object detection is crucial in diverse autonomous systems like surveillance, autonomous driving, and driver assistance, ensuring safety by recognizing pedestria ns, vehicles, traffic lights, and signs. However, adverse weather conditions such as snow, fog, and rain pose a challenge, affecting detection accuracy and risking accidents and damage. This clearly demonstrates the need for robust object detection solutions that work in all weather conditions. We employed three strategies to enhance deep learning-based object detection in adverse weather: training on real-world all-weather images, training on images with synthetic augmented weather noise, and integrating object detection with adverse weather image denoising. The synthetic weather noise is generated using analytical methods, GAN networks, and style-transfer networks. We compared the performance of these strategies by training object detection models using real-world all-weather images from the BDD100K dataset and for assessment employed unseen real-world adverse weather images. Adverse weather denoising methods were evaluated by denoising real-world adverse weather images and the results of object detection on denoised and original noisy images were compared. We found that the model trained using all-weather real-world images performed best, while the strategy of doing object detection on denoised images performed worst.

****

DiffBody: Diffusion-Based Pose and Shape Editing of Human Images

Yuta Okuyama, Yuki Endo, Yoshihiro Kanamori; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6333-6342

Pose and body shape editing in a human image has received increasing attention. However, current methods often struggle with dataset biases and deteriorate realism and the person's identity when users make large edits. We propose a one-shot approach that enables large edits with identity preservation. To enable large edits, we fit a 3D body model, project the input image onto the 3D model, and change the body's pose and shape. Because this initial textured body model has artifacts due to occlusion and the inaccurate body shape, the rendered image undergoes a diffusion-based refinement, in which strong noise destroys body structure and identity whereas insufficient noise does not help. We thus propose an iterative refinement with weak noise, applied first for the whole body and then for the face. We further enhance the realism by fine-tuning text embeddings via self-supervised learning. Our quantitative and qualitative evaluations demonstrate that our method outperforms other existing methods across various datasets.

****

Sound3DVDet: 3D Sound Source Detection Using Multiview Microphone Array and RGB Images

Yuhang He, Sangyun Shin, Anoop Cherian, Niki Trigoni, Andrew Markham; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5496-5507

Spatial localization of 3D sound sources is an important problem in many real world scenarios, especially when the sources may not have any visually distinguishable characteristics; e.g., finding a gas leak, a malfunctioning motor, etc. In this paper, we cast this task in a novel audio-visual setting, by introducing an acoustic-camera rig consisting of a centered pinhole RGB camera and an uniform circular array of four coplanar microphones. Using this setup, we propose Sound3DVDet - a 3D sound source localization Transformer model that takes as input the neural embeddings of the sound signals from the microphones and multiview images (with known poses), and learns to minimize the reprojection error between the predicted locations of the sound sources by the two modalities and the ground truth as the camera moves. When training to minimize this consistency loss, the model learns an implicit association between the audio heard at the microphones and the 3D spatial location in the RGB image, which is sufficient to localize the sources in 3D from a single RGB view. To evaluate our method, we introduce a new dataset: Sound3DVDet Dataset, consisting of nearly 6k scenes produced using the SoundSpaces simulator. We conduct extensive experiments on our dataset and shows the efficacy of our approach against closely related methods, demonstrating significant improvements in the localization accuracy.

****

Annotation-Free Audio-Visual Segmentation

Jinxiang Liu, Yu Wang, Chen Ju, Chaofan Ma, Ya Zhang, Weidi Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5604-5614

The objective of Audio-Visual Segmentation (AVS) is to localise the sounding objects within visual scenes by accurately predicting pixel-wise segmentation masks. To tackle the task, it involves a comprehensive consideration of both the data and model aspects. In this paper, first, we initiate a novel pipeline for generating artificial data for the AVS task without extra manual annotations. We leverage existing image segmentation and audio datasets and match the image-mask pairs with its corresponding audio samples using category labels in segmentation datasets, that allows us to effortlessly compose (image, audio, mask) triplets for training AVS models. The pipeline is annotation-free and scalable to cover a large number of categories. Additionally, we introduce a lightweight model SAMA-AVS which adapts the pre-trained segment anything model (SAM) to the AVS task. By introducing only a small number of trainable parameters with adapters, the proposed model can effectively achieve adequate audio-visual fusion and interaction in the encoding stage with vast majority of parameters fixed. We conduct extensive experiments, and the results show our proposed model remarkably surpasses other competing methods. Moreover, by using the proposed model pretrained with our synthetic data, the performance on real AVSBench data is further improved, achieving 83.17 mIoU on S4 subset and 66.95 mIoU on MS3 set. The project page is https://jinxiang-liu.github.io/anno-free-AVS/.

****

SC-MIL: Supervised Contrastive Multiple Instance Learning for Imbalanced Classification in Pathology

Dinkar Juyal, Siddhant Shingi, Syed Ashar Javed, Harshith Padigela, Chintan Shah, Anand Sampat, Archit Khosla, John Abel, Amaro Taylor-Weiner; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7946-7955

Multiple Instance learning (MIL) models have been extensively used in pathology to predict biomarkers and risk-stratify patients from gigapixel-sized images. Machine learning problems in medical imaging often deal with rare diseases, making it important for these models to work in a label-imbalanced setting. In pathology images, there is another level of imbalance, where given a positively labeled Whole Slide Image (WSI), only a fraction of pixels within it contribute to the positive label. This compounds the severity of imbalance and makes imbalanced classification in pathology challenging. Furthermore, these imbalances can occur in out-of-distribution (OOD) datasets when the models are deployed in the real-world. We leverage the idea that decoupling feature and classifier learning can lead to improved decision boundaries for label imbalanced datasets. To this end, we investigate the integration of supervised contrastive learning with multiple instance learning (SC-MIL). Specifically, we propose a joint-training MIL framework in the presence of label imbalance that progressively transitions from learning bag-level representations to optimal classifier learning. We perform experiments with different imbalance settings for two well-studied problems in cancer pathology: subtyping of non-small cell lung cancer and subtyping of renal cell carcinoma. SC-MIL provides large and consistent improvements over other techniques on both in-distribution (ID) and OOD held-out sets across multiple imbalanced settings.

****

MetaSeg: MetaFormer-Based Global Contexts-Aware Network for Efficient Semantic Segmentation

Beoungwoo Kang, Seunghun Moon, Yubin Cho, Hyunwoo Yu, Suk-Ju Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 434-443

Beyond the Transformer, it is important to explore how to exploit the capacity of the MetaFormer, an architecture that is fundamental to the performance improvements of the Transformer. Previous studies have exploited it only for the backbone network. Unlike previous studies, we explore the capacity of the Metaformer architecture more extensively in the semantic segmentation task. We propose a powerful semantic segmentation network, MetaSeg, which leverages the Metaformer arc hitecture from the backbone to the decoder. Our MetaSeg shows that the MetaFormer architecture plays a significant role in capturing the useful contexts for the decoder as well as for the backbone. In addition, recent segmentation methods have shown that using a CNN-based backbone for extracting the spatial information and a decoder for extracting the global information is more effective than using a transformer-based backbone with a CNN-based decoder. This motivates us to adopt the CNN-based backbone using the MetaFormer block and design our MetaFormer-based decoder, which consists of a novel self-attention module to capture the global contexts. To consider both the global contexts extraction and the computational efficiency of the self-attention for semantic segmentation, we propose a Channel Reduction Attention (CRA) module that reduces the channel dimension of the query and key into the one dimension. In this way, our proposed MetaSeg outperforms the previous state-of-the-art methods with more efficient computational costs on popular semantic segmentation and a medical image segmentation benchmark, including ADE20K, Cityscapes, COCO-stuff, and Synapse.

****

JOADAA: Joint Online Action Detection and Action Anticipation

Mohammed Guermal, Abid Ali, Rui Dai, François Brémond; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6889-6898

Action anticipation involves forecasting future actions by connecting the past events to future ones. However, this reasoning ignores the real-life hierarchy of events which is considered to be of three main parts: past, present, and future. We argue that considering these three main parts and their dependencies could improve performance. On the other hand, online action detection is the task of predicting actions in a streaming manner. In this case, one has access only to the past and present information. Therefore, in online action detection (OAD) the existing approaches miss semantics or future information which limits the performance of existing approaches. To sum up, for both of these tasks, the complete set of knowledge (past-present-future) is missing, which makes it challenging to infer action dependencies achieving good performances. To address this limitation, we propose fusing both tasks in one uniform architecture. By combining action anticipation and online action detection, our approach can cover the missing dependencies of future information in online action detection. This method, referred as JOADAA, presents a uniform model that jointly performs action anticipation and online action detection. We validate our proposed model on three challenging datasets: THUMOS, which is a sparsely annotated dataset with one action per time step, CHARADES and Multi-THUMOS, two densely annotated datasets, with more complex scenarios. JOADAA achieves SOTA results on these benchmarks for both tasks.

****

Denoising and Selecting Pseudo-Heatmaps for Semi-Supervised Human Pose Estimation

Zhuoran Yu, Manchen Wang, Yanbei Chen, Paolo Favaro, Davide Modolo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6280-6289

We propose a new semi-supervised learning design for human pose estimation that revisits the popular dual-student framework and enhances it two ways. First, we introduce a denoising scheme to generate reliable pseudo-heatmaps as targets for learning from unlabeled data. This uses multi-view augmentations and a threshold-and-refine procedure to produce a pool of pseudo-heatmaps. Second, we select the learning targets from these pseudo-heatmaps guided by the estimated cross-student uncertainty. We evaluate our proposed method on multiple evaluation setups on the COCO benchmark. Our results show that our model outperforms previous state-of-the-art semi-supervised pose estimators, especially in extreme low-data regime. For example with only 0.5K labeled images our method is capable of surpassing the best competitor by 7.22 mAP (+25% absolute improvement). We also demonstrate that our model can learn effectively from unlabeled data in the wild to further boost its generalization and performance.

****

CSAM: A 2.5D Cross-Slice Attention Module for Anisotropic Volumetric Medical Image Segmentation

Alex Ling Yu Hung, Haoxin Zheng, Kai Zhao, Xiaoxi Du, Kaifeng Pang, Qi Miao, Steven S. Raman, Demetri Terzopoulos, Kyunghyun Sung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5923-5932

A large portion of volumetric medical data, especially magnetic resonance imaging (MRI) data, is anisotropic, as the through-plane resolution is typically much lower than the in-plane resolution. Both 3D and purely 2D deep learning-based segmentation methods are deficient in dealing with such volumetric data since the performance of 3D methods suffers when confronting anisotropic data, and 2D methods disregard crucial volumetric information. Insufficient work has been done on 2.5D methods, in which 2D convolution is mainly used in concert with volumetric information. These models focus on learning the relationship across slices, but typically have many parameters to train. We offer a Cross-Slice Attention Module (CSAM) with minimal trainable parameters, which captures information across all the slices in the volume by applying semantic, positional, and slice attention on deep feature maps at different scales. Our extensive experiments using different network architectures and tasks demonstrate the usefulness and generalizability of CSAM. Associated code is available at https://github.com/aL3x-O-o-Hung/CSAM.

****

Segment Anything, From Space?

Simiao Ren, Francesco Luzi, Saad Lahrichi, Kaleb Kassaw, Leslie M. Collins, Kyle Bradbury, Jordan M. Malof; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8355-8365

Recently, the first foundation model developed specifically for image segmentation tasks was developed, termed the "Segment Anything Model" (SAM). SAM can segment objects in input imagery based on cheap input prompts, such as one (or more) points, a bounding box, or a mask. The authors examined the zero-shot image segmentation accuracy of SAM on a large number of vision benchmark tasks and found that SAM usually achieved recognition accuracy similar to, or sometimes exceeding, vision models that had been trained on the target tasks. The impressive generalization of SAM for segmentation has major implications for vision researchers working on natural imagery. In this work, we examine whether SAM's performance extends to overhead imagery problems and help guide the community's response to its development. We examine SAM's performance on a set of diverse and widely studied benchmark tasks. We find that SAM does often generalize well to overhead imagery, although it fails in some cases due to the unique characteristics of overhead imagery and its common target objects. We report on these unique systematic failure cases for remote sensing imagery that may comprise useful future research for the community.

****

UOW-Vessel: A Benchmark Dataset of High-Resolution Optical Satellite Images for Vessel Detection and Segmentation

Ly Bui, Son Lam Phung, Yang Di, Thanh Le, Tran Thanh Phong Nguyen, Sandy Burden, Abdesselam Bouzerdoum; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4428-4436

In this paper, we introduce UOW-Vessel, a benchmark dataset of high-resolution optical satellite images for vessel detection and segmentation. Our dataset consists of 3,500 images, collected from 14 countries across 4 continents. With a total of 35,598 instances in 10 vessel categories, UOW-Vessel is to date the largest satellite image dataset for vessel recognition. Furthermore, compared to the existing public datasets that only provide bounding box ground-truth, our new dataset offers more accurate polygon annotations of vessel objects. This dataset is expected to support instance segmentation-based approaches, which is a less investigated area in vessel surveillance. We also report extensive evaluations of the recent algorithms for instance segmentation on the new benchmark dataset.

****

Single Frame Semantic Segmentation Using Multi-Modal Spherical Images

Suresh Guttikonda, Jason Rambach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3222-3231

In recent years, the research community has shown a lot of interest to panoramic images that offer a 360-degree directional perspective. Multiple data modalities can be fed, and complimentary characteristics can be utilized for more robust and rich scene interpretation based on semantic segmentation, to fully realize the potential. Existing research, however, mostly concentrated on pinhole RGB-X semantic segmentation. In this study, we propose a transformer-based cross-modal fusion architecture to bridge the gap between multi-modal fusion and omnidirectional scene perception. We employ distortion-aware modules to address extreme object deformations and panorama distortions that result from equirectangular representation. Additionally, we conduct cross-modal interactions for feature rectification and information exchange before merging the features in order to communicate long-range contexts for bi-modal and tri-modal feature streams. In thorough tests using combinations of four different modality types in three indoor panoramic-view datasets, our technique achieved state-of-the-art mIoU performance: 60.60% on Stanford2D3DS (RGB-HHA), 71.97% on Structured3D (RGB-D-N), and 35.92% on Matterport3D (RGB-D).

****

SynthProv: Interpretable Framework for Profiling Identity Leakage

Jaisidh Singh, Harshil Bhatia, Mayank Vatsa, Richa Singh, Aparna Bharati; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4746-4756

Generative Adversarial Networks (GANs) can generate hyperrealistic face images of synthetic identities based on a latent understanding of real images from a large training set. Despite their proficiency, the term "synthetic identity" remains ambiguous, and the uniqueness of the faces GANs produce is rarely assessed. Recent studies have found that identities from the training data can unintentionally appear in the faces generated by StyleGAN2, but the cause of this phenomenon is unclear. In this work, we propose a novel framework, SynthProv, that utilizes the improved interpolation ability of StyleGAN2 latent space and employs image composition to analyze leakage. This is the first method that goes beyond detection and traces the source or provenance of constituent identity signals in the generated image. Experiments show that SynthProv succeeds in both detection and provenance tasks using multiple matching strategies. We identify identities from FFHQ and CelebA-HQ training datasets with the highest leakage into the latent space as "leaking reals". Analyzing latent space behavior to evaluate generative model privacy via leakage is an important research direction, as undetected leaking reals pose a significant threat to training data privacy. Our code is available at https://github.com/jaisidhsingh/SynthProv

****

Discovering and Mitigating Biases in CLIP-Based Image Editing

Md Mehrab Tanjim, Krishna Kumar Singh, Kushal Kafle, Ritwik Sinha, Garrison W. Cottrell; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2984-2993

In recent years, the use of CLIP (Contrastive Language-Image Pre-Training) has become increasingly popular in a wide range of downstream applications, including zero-shot image classification and text-to-image synthesis. Despite being trained on a vast dataset, the CLIP model has been found to exhibit biases against certain protected attributes, such as gender and race. While previous research has focused on the impact of such biases on image classification, there has been little investigation into their effects on CLIP-based generative tasks. In this paper, we aim to address this gap in the literature by uncovering the queries for which the CLIP model introduces biases in the text-based image editing task. Through a series of experiments, we demonstrate that these biases can have a significant impact on the quality and content of the generated images. To mitigate these biases, we propose a debiasing technique that does not require retraining either the CLIP model or the underlying generative model. Our results show that our proposed framework can effectively reduce the impact of biases in CLIP-based image editing models. Overall, this paper highlights the importance of addressing biases in CLIP-based generative tasks and provides practical solutions that can be readily adopted by researchers and practitioners working in this area.

****

Repetitive Action Counting With Motion Feature Learning

Xinjie Li, Huijuan Xu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6499-6508

Repetitive action counting aims to count the number of repetitive actions in a video. The critical challenge of this task is to uncover the periodic pattern between repetitive actions by computing feature similarity between frames. However, existing methods only rely on the RGB feature of each frame to compute the feature similarity while neglecting the background change of repetitive actions. The abrupt background change may cause feature discrepancies of the same action moment and lead to errors in counting. To this end, we propose a two-branch framework, i.e., RGB and motion branches, with the motion branch complementing the RGB branch to enhance the foreground motion feature learning. Specifically, foreground motion features are highlighted with flow-guided attention on frame features. In addition, to alleviate the noise from moving background distractors and reinforce the periodic pattern, we propose a temporal self-similarity matrix reconstruction loss to improve the temporal correspondence between the same motion feature from different frames. Lastly, to make the motion feature effectively supplement the RGB feature, we present a novel variance-prompted loss weights generation technique to automatically generate dynamic loss weights for two branches in collaborative training. Extensive experiments are conducted on the RepCount and UCFRep datasets to verify our proposed method with state-of-the-art performance. Our method also achieves the best performance on the cross-dataset generalization experiment.

****

Deep Image Fingerprint: Towards Low Budget Synthetic Image Detection and Model Lineage Analysis

Sergey Sinitsa, Ohad Fried; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4067-4076

The generation of high-quality images has become widely accessible and is a rapidly evolving process. As a result, anyone can generate images that are indistinguishable from real ones. This leads to a wide range of applications, including malicious usage with deceptive intentions. Despite advances in detection techniques for generated images, a robust detection method still eludes us. Furthermore, model personalization techniques might affect the detection capabilities of existing methods. In this work, we utilize the architectural properties of convolutional neural networks (CNNs) to develop a new detection method. Our method can detect images from a known generative model and enable us to establish relationships between fine-tuned generative models. We tested the method on images produced by both Generative Adversarial Networks (GANs) and recent large text-to-image models (LTIMs) that rely on Diffusion Models. Our approach outperforms others trained under identical conditions and achieves comparable performance to state-of-the-art pre-trained detection methods on images generated by Stable Diffusion and MidJourney, with significantly fewer required train samples.

****

HALSIE: Hybrid Approach to Learning Segmentation by Simultaneously Exploiting Image and Event Modalities

Shristi Das Biswas, Adarsh Kosta, Chamika Liyanagedera, Marco Apolinario, Kaushik Roy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5964-5974

Event cameras detect changes in per-pixel intensity to generate asynchronous 'event streams'. They offer great potential for accurate semantic map retrieval in real-time autonomous systems owing to their much higher temporal resolution and high dynamic range (HDR) compared to conventional cameras. However, existing implementations for event-based segmentation suffer from sub-optimal performance since these temporally dense events only measure the varying component of a visual signal, limiting their ability to encode dense spatial context compared to frames. To address this issue, we propose a hybrid end-to-end learning framework HALSIE, utilizing three key concepts to reduce inference cost by up to 20x versus p rior art while retaining similar performance: First, a simple and efficient cross-domain learning scheme to extract complementary spatio-temporal embeddings from both frames and events. Second, a specially designed dual-encoder scheme with Spiking Neural Network (SNN) and Artificial Neural Network (ANN) branches to minimize latency while retaining cross-domain feature aggregation. Third, a multi-scale cue mixer to model rich representations of the fused embeddings. These qualities of HALSIE allow for a very lightweight architecture achieving state-of-the-art segmentation performance on DDD-17, MVSEC, and DSEC-Semantic datasets with up to 33x higher parameter efficiency and favorable inference cost (17.9mJ per cycle). Our ablation study also brings new insights into effective design choices that can prove beneficial for research across other vision tasks.

****

Hierarchical Diffusion Autoencoders and Disentangled Image Manipulation

Zeyu Lu, Chengyue Wu, Xinyuan Chen, Yaohui Wang, Lei Bai, Yu Qiao, Xihui Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5374-5383

Diffusion models have attained impressive visual quality for image synthesis. However, how to interpret and manipulate the latent space of diffusion models has not been extensively explored. Prior work diffusion autoencoders encode the semantic representations into a semantic latent code, which fails to reflect the rich information of details and the intrinsic feature hierarchy. To mitigate those limitations, we propose Hierarchical Diffusion Autoencoders (HDAE) that exploit the fine-grained-to-abstract and low-level-to-high-level feature hierarchy for the latent space of diffusion models. The hierarchical latent space of HDAE inherently encodes different abstract levels of semantics and provides more comprehensive semantic representations. In addition, we propose a truncated-feature-based approach for disentangled image manipulation. We demonstrate the effectiveness of our proposed approach with extensive experiments and applications on image reconstruction, style mixing, controllable interpolation, detail-preserving and disentangled image manipulation, and multi-modal semantic image synthesis.

****

Improved Topological Preservation in 3D Axon Segmentation and Centerline Detection Using Geometric Assessment-Driven Topological Smoothing (GATS)

Nina I. Shamsi, Alec S. Xu, Lars A. Gjesteby, Laura J. Brattain; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8005-8014

Automated axon tracing via fully supervised learning requires large amounts of 3D brain imagery, which is time consuming and laborious to obtain. It also requires expertise. Thus, there is a need for more efficient segmentation and centerline detection techniques to use in conjunction with automated annotation tools. Topology-preserving methods ensure that segmented components maintain geometric connectivity, which is especially meaningful for applications where volumetric data is used, and these methods often make use of morphological thinning algorithms as the thinned outputs can be useful for both segmentation and centerline detection of curvilinear structures. Current morphological thinning approaches used in conjunction with topology-preserving methods are prone to over-thinning and require manual configuration of hyperparameters. We propose an automated approach for morphological smoothing using geometric assessment of the radius of tubular structures in brain microscopy volumes, and apply average pooling to prevent over-thinning. We use this approach to formulate a loss function, which we call Geometric Assessment-driven Topological Smoothing loss, or GATS. Our approach increased segmentation and centerline detection evaluation metrics by 2%-5% across multiple datasets, and improved the Betti error rates by 9%. Our ablation study showed that geometric assessment of tubular structures achieved higher segmentation and centerline detection scores, and using average pooling for morphological smoothing in place of thinning algorithms reduced the Betti errors. We observed increased topological preservation during automated annotation of 3D axons volumes from models trained with GATS.

****

CycleCL: Self-Supervised Learning for Periodic Videos

Matteo Destro, Michael Gygli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2861-2870

Analyzing periodic video sequences is a key topic in applications such as automatic production systems, remote sensing, medical applications, or physical training. An example is counting repetitions of a physical exercise. Due to the distinct characteristics of periodic data, self-supervised methods designed for standard image datasets do not capture changes relevant to the progression of the cycle and fail to ignore unrelated noise. They thus do not work well on periodic data. In this paper, we propose CycleCL, a self-supervised learning method specifically designed to work with periodic data. We start from the insight that a good visual representation for periodic data should be sensitive to the phase of a cycle, but be invariant to the exact repetition, i.e. it should generate identical representations for a specific phase throughout all repetitions. We exploit the repetitions in videos to design a novel contrastive learning method based on a triplet loss that optimizes for these desired properties. Our method uses pre-trained features to sample pairs of frames from approximately the same phase and negative pairs of frames from different phases. Then, we iterate between optimizing a feature encoder and resampling triplets, until convergence. By optimizing a model this way, we are able to learn features that have the mentioned desired properties. We evaluate CycleCL on an industrial and multiple human actions datasets, where it significantly outperforms previous video-based self-supervised learning methods on all tasks.

****

DR10K: Transfer Learning Using Weak Labels for Grading Diabetic Retinopathy on DR10K Dataset

Mohamed ElHabebe, Shereen ElKordi, Ahmed Gamal ElDin, Noha Adly, Marwan Torki, Ahmed Elmassry, Islam SH Ahmed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7748-7758

In this paper, we contrast the usage of two deep-learning approaches for the automatic grading of diabetic retinopathy (DR) and diabetic macular edema (DME) in retinal fundus photographs using a relatively small novel dataset. We developed a telemedicine system to collect and humanly grade 11,109 diabetic patients. The certified graders annotated the level of DR as well as the existence of a referable DME in the macula-centered fundus images only. We use EfficientNet to build an AI-based model for both problems. To examine the transfer learning validity, the model was trained on an external dataset (EyePacs) and then finetuned on the egyptian data for the DR and DME grading problems. Firstly, we use the macula-centered images only in fine-tuning. Secondly, we use optic-disc-centered images in addition to macula-centered images. We obtained the labels for the optic-disc-centered images directly from the corresponding macula-centered labels as weak labels. Then, both types of images are used in fine-tuning. We found an increase in the DR performance using the second approach in both accuracy and quadratic weighted kappa(QWK). Notably, QWK increased from 90.23% to 91.3% using additional weakly labeled optic-disc-centered fundus images.

****

Open-NeRF: Towards Open Vocabulary NeRF Decomposition

Hao Zhang, Fang Li, Narendra Ahuja; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3456-3465

In this paper, we address the challenge of decomposing Neural Radiance Fields (NeRF) into objects from an open vocabulary, a critical task for object manipulation in 3D reconstruction and view synthesis. Current techniques for NeRF decomposition involve a trade-off between the flexibility of processing open-vocabulary queries and the accuracy of 3D segmentation. We present, Open-vocabulary Embedded Neural Radiance Fields (Open-NeRF), that leverage large-scale, off-the-shelf, segmentation models like the Segment Anything Model (SAM) and introduce an integrate-and-distill paradigm with hierarchical embeddings to achieve both the flexibility of open-vocabulary querying and 3D segmentation accuracy. Open-NeRF first utilizes large-scale foundation models to generate hierarchical 2D mask proposals from varying viewpoints. These proposals are then aligned via tracking approaches and integrated within the 3D space and subsequently distilled into the 3D f ield. This process ensures consistent recognition and granularity of objects from different viewpoints, even in challenging scenarios involving occlusion and indistinct features. Our experimental results show that the proposed Open-NeRF outperforms state-of-the-art methods such as LERF and FFD in open-vocabulary scenarios. Open-NeRF offers a promising solution to NeRF decomposition, guided by open-vocabulary queries, enabling novel applications in robotics and vision-language interaction in open-world 3D scenes. Please find the code at https://github.com/haoz19/Open-NeRF

****

Exploiting the Signal-Leak Bias in Diffusion Models

Martin Nicolas Everaert, Athanasios Fitsios, Marco Bocchio, Sami Arpa, Sabine Süsstrunk, Radhakrishna Achanta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4025-4034

There is a bias in the inference pipeline of most diffusion models. This bias arises from a signal leak whose distribution deviates from the noise distribution, creating a discrepancy between training and inference processes. We demonstrate that this signal-leak bias is particularly significant when models are tuned to a specific style, causing sub-optimal style matching. Recent research tries to avoid the signal leakage during training. We instead show how we can exploit this signal-leak bias in existing diffusion models to allow more control over the generated images. This enables us to generate images with more varied brightness, and images that better match a desired style or color. By modeling the distribution of the signal leak in the spatial frequency and pixel domains, and including a signal leak in the initial latent, we generate images that better match expected results without any additional training.

****

Weakly-Supervised Representation Learning for Video Alignment and Analysis

Guy Bar-Shalom, George Leifman, Michael Elad; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6909-6919

Many tasks in video analysis and understanding boil down to the need for frame-based feature learning, aiming to encapsulate the relevant visual content so as to enable simpler and easier subsequent processing. While supervised strategies for this learning task can be envisioned, self and weakly-supervised alternatives are preferred due to the difficulties in getting labeled data. This paper introduces LRProp -- a novel weakly-supervised representation learning approach, with an emphasis on the application of temporal alignment between pairs of videos of the same action category. The proposed approach uses a transformer encoder for extracting frame-level features, and employs the DTW algorithm within the training iterations in order to identify the alignment path between video pairs. Through a process referred to as "pair-wise position propagation", the probability distributions of these correspondences per location are matched with the similarity of the frame-level features via KL-divergence minimization. The proposed algorithm uses also a regularized SoftDTW loss for better tuning the learned features. Our novel representation learning paradigm consistently outperforms the state of the art on temporal alignment tasks, establishing a new performance bar over several downstream video analysis applications.

****

NCIS: Neural Contextual Iterative Smoothing for Purifying Adversarial Perturbations

Sungmin Cha, Naeun Ko, Heewoong Choi, Youngjoon Yoo, Taesup Moon; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3789-3799

We propose a novel and effective purification-based adversarial defense method against pre-processor blind white- and black-box attacks, without requiring any adversarial training or retraining of the classification model. Based on the observation of the adversarial noise, we propose a simple iterative Gaussian Smoothing (GS) that smoothes out adversarial noise and achieves substantially high robust accuracy. To further improve the method, we propose Neural Contextual Iterative Smoothing (NCIS), which trains a blind-spot network (BSN) in a self-supervised manner to reconstruct the discriminative features of the smoothed original ima

ge. From the extensive experiments on the large-scale ImageNet, we show that our method achieves both competitive standard accuracy and state-of-the-art robust accuracy against most strong purifier-blind white- and black-box attacks. Also, we propose a new evaluation benchmark based on commercial image classification APIs, including AWS, Azure, Clarifai, and Google, and demonstrate that users can use our method to increase the adversarial robustness of APIs.
****
D3GU: Multi-Target Active Domain Adaptation via Enhancing Domain Alignment
Lin Zhang, Linghan Xu, Saman Motamed, Shayok Chakraborty, Fernando De la Torre; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2577-2586
Unsupervised domain adaptation (UDA) for image classification has made remarkable progress in transferring classification knowledge from a labeled source domain to an unlabeled target domain, thanks to effective domain alignment techniques. Recently, in order to further improve performance on a target domain, many Single-Target Active Domain Adaptation (ST-ADA) methods have been proposed to identify and annotate the salient and exemplar target samples. However, it requires one model to be trained and deployed for each target domain and the domain label associated with each test sample. This largely restricts its application in the ubiquitous scenarios with multiple target domains. Therefore, we propose a Multi-Target Active Domain Adaptation (MT-ADA) framework for image classification, named D3GU, to simultaneously align different domains and actively select samples from them for annotation. This is the first research effort in this field to our best knowledge. D3GU applies Decomposed Domain Discrimination (D3) during training to achieve both source-target and target-target domain alignments. Then during active sampling, a Gradient Utility (GU) score is designed to weight every unlabeled target image by its contribution towards classification and domain alignment tasks, and is further combined with KMeans clustering to form GU-KMeans for diverse image sampling. Extensive experiments on three benchmark datasets, Office31, OfficeHome, and DomainNet, have been conducted to validate consistently superior performance of D3GU for MT-ADA.
****
Fixed Pattern Noise Removal for Multi-View Single-Sensor Infrared Camera
Arnaud Barral, Pablo Arias, Axel Davy; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1669-1678
Fixed pattern noise (FPN) is a temporally coherent noise present on videos due to the non-uniformities in the response of the imaging sensor. It is a common problem for infrared videos which degrades the quality of the observation and hinders subsequent applications. In this work we introduce a generalization of the FPN removal problem where the input data consists of several different sequences with the same FPN. This is motivated by infrared cameras that capture multiple views with a single sensor via a periodic motion pattern of a mirror or the camera itself, such as those used in surveillance. This multi-view setting allows for a much more accurate estimation of the FPN in comparison with the standard FPN removal problem from a single view. We propose a novel energy minimization approach for multi-view FPN removal, and two optimization algorithms that can be applied both in an off-line and on-line manner. In addition, we show that the proposed energy can be adapted to the problem of FPN removal from a single view with a rolling window approach, obtaining a significant improvement over the state of the art. We demonstrate the performance of the proposed method with synthetic data and real data from surveillance infrared cameras.
****
Painterly Image Harmonization via Adversarial Residual Learning
Xudong Wang, Li Niu, Junyan Cao, Yan Hong, Liqing Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5141-5150
Image compositing plays a vital role in photo editing. After inserting a foreground object into another background image, the composite image may look unnatural and inharmonious. When the foreground is photorealistic and the background is an artistic painting, painterly image harmonization aims to transfer the style of

background painting to the foreground object, which is a challenging task due to the large domain gap between foreground and background. In this work, we employ adversarial learning to bridge the domain gap between foreground feature map and background feature map. Specifically, we design a dual-encoder generator, in which the residual encoder produces the residual features added to the foreground feature map from main encoder. Then, a pixel-wise discriminator plays against the generator, encouraging the refined foreground feature map to be indistinguishable from background feature map. Extensive experiments demonstrate that our method could achieve more harmonious and visually appealing results than previous methods.

****

CoD: Coherent Detection of Entities From Images With Multiple Modalities

Vinay Verma, Dween Sanny, Abhishek Singh, Deepak Gupta; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8015-8024

However, in real-world scenarios, multiple sources of data in different modalities are often present, making it difficult to accurately define object boundaries for various products or information. For instance, while extracting information from a document, it may be necessary to utilize both visual information (e.g., image/object) and textual information from OCR to detect and classify information associated with objects, such as text blocks, tables, and figures. If visual and textual information pertain to the same object, the model should detect the bounding box around all multi-modal information. The problem of object detection in computer vision has traditionally been viewed as a unimodal problem in the literature, which poses a significant challenge. This work presents a novel approach to automating object boundary identification in multi-modal scenarios. The study proposes an end-to-end method that employs transformers for detecting object boundaries in a multi-modal environment. The proposed model takes multi-scale image features, OCR-based text extraction, and 2D position embedding of words as input, which interact through self- and cross-attention mechanisms. Additionally, the study proposes a domain adaptation model to address the often significant domain gap between training and test samples in such scenarios. The proposed approach shows a significant improvement of 27.2%, 5.0% and 1.7% using hard negative samples, multi-modal and domain shift scenarios, respectively. The ablation studies confirm the effectiveness of the proposed components.

****

Improving Graph Networks Through Selection-Based Convolution

David Hart, Bryan Morse; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1794-1804

Graph Convolutional Networks (GCNs) provide a general framework that can learn in a variety of data domains, such as 3D geometry, social networks, and chemical structures. GCNs, however, often ignore intrinsic relationships among nodes in the graph, and these relationships need to be learned indirectly during the training process through mechanisms such as attention or local-kernel approximation. This paper introduces selection-based graph convolution, a method for preserving these intrinsic relationships within the graph convolution operator which provides improved performance over attention-based counterparts on various tasks. We demonstrate the effectiveness of selection to improve the performance of many types of GCNs on tasks such as spatial graph classification. Furthermore, we demonstrate the ability to improve state-of-the-art graph networks for road traffic estimation and molecular property prediction.

****

Beyond Fusion: Modality Hallucination-Based Multispectral Fusion for Pedestrian Detection

Qian Xie, Ta-Ying Cheng, Jia-Xing Zhong, Kaichen Zhou, Andrew Markham, Niki Trigoni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 655-664

Pedestrian detection is a fundamental task for many downstream applications. Visible and thermal images, as the two most important data types, are usually used to detect pedestrians under various environmental conditions. Many state-of-the-art works have been proposed to use two-stream (i.e., two-branch) architectures to combine visible and thermal information to improve detection performance. However, conventional visible-thermal fusion-based methods have no ability to obtain useful information from the visible branch under poor visibility conditions. The visible branch could even sometimes bring noise into the combined features. In this paper, we present a novel thermal and visible fusion architecture for pedestrian detection. Instead of simply using two branches to separately extract thermal and visible features and then fusing them, we introduce a hallucination branch to learn the mapping from thermal to visible domain, forming a three-branch feature extraction module. We then adaptively fuse feature maps from all the three branches (i.e., thermal, visible, and hallucination). With this new integrated hallucination branch, our network can still get relatively good visible feature maps under challenging low visibility conditions, thus boosting the overall detection performance. Finally, we experimentally demonstrate the superiority of the proposed architecture over conventional fusion methods.

****

BSRAW: Improving Blind RAW Image Super-Resolution

Marcos V. Conde, Florin Vasluianu, Radu Timofte; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8500-8510

In smartphones and compact cameras, the Image Signal Processor (ISP) transforms the RAW sensor image into a human-readable sRGB image. Most popular super-resolution methods depart from a sRGB image and upscale it further, improving its quality. However, modeling the degradations in the sRGB domain is complicated because of the non-linear ISP transformations. Despite this known issue, only a few methods work directly with RAW images and tackle real-world sensor degradations. We tackle blind image super-resolution in the RAW domain. We design a realistic degradation pipeline tailored specifically for training models with raw sensor data. Our approach considers sensor noise, defocus, exposure, and other common issues. Our BSRAW models trained with our pipeline can upscale real-scene RAW images and improve their quality. As part of this effort, we also present a new DSLM dataset and benchmark for this task.

****

SICKLE: A Multi-Sensor Satellite Imagery Dataset Annotated With Multiple Key Cropping Parameters

Depanshu Sani, Sandeep Mahato, Sourabh Saini, Harsh Kumar Agarwal, Charu Chandra Devshali, Saket Anand, Gaurav Arora, Thiagarajan Jayaraman; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5995-6004

The availability of well-curated datasets has driven the success of Machine Learning (ML) models. Despite greater access to earth observation data in agriculture, there is a scarcity of curated and labelled datasets, which limits the potential of its use in training ML models for remote sensing (RS) in agriculture. To this end, we introduce a first-of-its-kind dataset called SICKLE, which constitutes a time-series of multi-resolution imagery from 3 distinct satellites: Landsat-8, Sentinel-1 and Sentinel-2. Our dataset constitutes multi-spectral, thermal and microwave sensors during January 2018 - March 2021 period. We construct each temporal sequence by considering the cropping practices followed by farmers primarily engaged in paddy cultivation in the Cauvery Delta region of Tamil Nadu, India; and annotate the corresponding imagery with key cropping parameters at multiple resolutions (i.e. 3m, 10m and 30m). Our dataset comprises 2, 370 season-wise samples from 388 unique plots, having an average size of 0.38 acres, for classifying 21 crop types across 4 districts in the Delta, which amounts to approximately 209, 000 satellite images. Out of the 2, 370 samples, 351 paddy samples from 145 plots are annotated with multiple crop parameters; such as the variety of paddy, its growing season and productivity in terms of per-acre yields. Ours is also one among the first studies that consider the growing season activities pertinent to crop phenology (spans sowing, transplanting and harvesting dates) as parameters of interest. We benchmark SICKLE on three tasks: crop type, crop phenology (sowing, transplanting, harvesting), and yield prediction.

****

LatentPaint: Image Inpainting in Latent Space With Diffusion Models

Ciprian Corneanu, Raghudeep Gadde, Aleix M. Martinez; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4334-4343

Image inpainting is generally done using either a domain-specific (preconditioned) model or a generic model that is postconditioned at inference time. Preconditioned models are fast at inference time but extremely costly to train, requiring training on each domain they are applied to. Postconditioned models do not require any domain-specific training but are slow during inference, requiring multiple forward and backward passes to converge to a desirable solution. Here, we derive an approach that does not require any domain specific training, yet is fast at inference time. To solve the costly inference computational time, we perform the forward-backward fusion step on a latent space rather than the image space. This is solved with a newly proposed propagation module in the diffusion process. Experiments on a number of domains demonstrate our approach attains or improves state-of-the-art results with the advantages of preconditioned and postconditioned models and none of their disadvantages.

****

Efficient Semantic Matching With Hypercolumn Correlation

Seungwook Kim, Juhong Min, Minsu Cho; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 139-148

Recent studies show that leveraging the match-wise relationships within the 4D correlation map yields significant improvements in establishing semantic correspondences - but at the cost of increased computation and latency. In this work, we focus on the aspect that the performance improvements of recent methods can also largely be attributed to the usage of multi-scale correlation maps, which hold various information ranging from low-level geometric cues to high-level semantic contexts. To this end, we propose HCCNet, an efficient yet effective semantic matching method which exploits the full potential of multi-scale correlation maps, while eschewing the reliance on expensive match-wise relationship mining on the 4D correlation map. Specifically, HCCNet performs feature slicing on the bottleneck features to yield a richer set of intermediate features, which are used to construct a hypercolumn correlation. HCCNet can consequently establish semantic correspondences in an effective manner by reducing the volume of conventional high-dimensional convolution or self-attention operations to efficient point-wise convolutions. HCCNet demonstrates state-of-the-art or competitive performances on the standard benchmarks of semantic matching, while incurring a notably lower latency and computation overhead compared to the existing SoTA methods.

****

OptFlow: Fast Optimization-Based Scene Flow Estimation Without Supervision

Rahul Ahuja, Chris Baker, Wilko Schwarting; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3161-3170

Scene flow estimation is a crucial component in the development of autonomous driving and 3D robotics, providing valuable information for environment perception and navigation. Despite the advantages of learning-based scene flow estimation techniques, their domain specificity and limited generalizability across varied scenarios pose challenges. In contrast, non-learning optimization-based methods, incorporating robust priors or regularization, offer competitive scene flow estimation performance, require no training, and show extensive applicability across datasets, but suffer from lengthy inference times. In this paper, we present OptFlow, a fast optimization-based scene flow estimation method. Without relying on learning or any labeled datasets, OptFlow achieves state-of-the-art performance for scene flow estimation on popular autonomous driving benchmarks. It integrates a local correlation weight matrix for correspondence matching, an adaptive correspondence threshold limit for nearest-neighbor search, and graph prior rigidity constraints, resulting in expedited convergence and improved point correspondence identification. Moreover, we demonstrate how integrating a point cloud registration function within our objective function bolsters accuracy and differentiates between static and dynamic points without relying on external odometry data. Consequently, OptFlow outperforms the baseline graph-prior method by approxi

mately 20% and the Neural Scene Flow Prior method by 5%-7% in accuracy, all while offering the fastest inference time among all non-learning scene flow estimation methods.
****

HashReID: Dynamic Network With Binary Codes for Efficient Person Re-Identification
Kshitij Nikhal, Yujunrong Ma, Shuvra S. Bhattacharyya, Benjamin S. Riggan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6046-6055
Biometric applications, such as person re-identification (ReID), are often deployed on energy constrained devices. While recent ReID methods prioritize high retrieval performance, they often come with large computational costs and high search time, rendering them less practical in real-world settings. In this work, we propose an input-adaptive network with multiple exit blocks, that can terminate computation early if the retrieval is straightforward or noisy, saving a lot of computation. To assess the complexity of the input, we introduce a temporal-based classifier driven by a new training strategy. Furthermore, we adopt a binary hash code generation approach instead of relying on continuous-valued features, which significantly improves the search process by a factor of 20. To ensure similarity preservation, we utilize a new ranking regularizer that bridges the gap between continuous and binary features. Extensive analysis of our proposed method is conducted on three datasets: Market1501, MSMT17 (Multi-Scene Multi-Time), and the BGC1 (BRIAR Government Collection). Using our approach, more than 70% of the samples with compact hash codes exit early on the Market1501 dataset, saving 80% of the networks computational cost and improving over other hash-based methods by 60%. These results demonstrate a significant improvement over dynamic networks and showcase comparable accuracy performance to conventional ReID methods.
****

Conditional Velocity Score Estimation for Image Restoration
Ziqiang Shi, Rujie Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 179-188
This paper proposes a new image restoration method by introducing a velocity variable on top of the data position during recovery. Under the guidance of the degraded image, it can effectively and dynamically control the direction of the diffusion path in the reverse-time stochastic differential equation (SDE). So the crucial factor is how to combine the degraded signal as a guide in this second-order reverse process with velocity, especially in the moving direction as a diffusion path. To this end, we propose a conditional velocity score approximation (CVSA) method based on the Bayesian principle to approximate the true posterior conditional velocity score by the sum of a priori conditional velocity score and an observation velocity score of the degraded measurement at the current moment. Our method is versatile from two perspectives. It can be used for both non-blind restoration and blind restoration. At the same time, there is almost no requirement for the degradation operator, and both linear and nonlinear tasks are acceptable. In non-blind restoration, including deblurring, inpainting, super-resolution, phase retrieval, and blind restoration, such as deblurring experiments, CVSA is better than other methods and achieves a new state-of-the-art.
****

Stochastic Binary Network for Universal Domain Adaptation
Saurabh Kumar Jain, Sukhendu Das; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 107-116
Universal domain adaptation (UniDA) is the unsupervised domain adaptation with label shift. UniDA aims to classify unlabeled target samples into one of the "known" categories or into a single "unknown" category. Its main challenge lies in detecting private classes from both domains and performing alignment between the common classes. Current methods employ various techniques and loss functions to address these challenges. However, these methods commonly represent classifiers as point weight vectors, which are prone to overfitting by the source domain samples due to the lack of supervision from the target domain. Consequently, these classifiers struggle to separate target samples into known and unknown categorie

s effectively. To address this, we introduce a novel framework called Stochastic Binary Network for Universal Domain Adaptation (STUN). STUN uses a Stochastic binary classifier for each class, whose weight is modeled as Gaussian distribution, enabling to sample an arbitrary number of classifiers while keeping the model size same as of two classifiers. Consistency between these sampled classifiers is used to derive the confidence scores for both source and target samples, which facilitates the alignment of common classes using weighted adversarial learning. Finally, we use deep discriminative clustering to formulate a loss function for solving the problem of fragmented feature distributions in the target domain. Extensive ablation studies and state-of-the-art results across three standard benchmark datasets show the efficacy of our framework.

****

FG-Net: Facial Action Unit Detection With Generalizable Pyramidal Features

Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, Mohammad Soleymani; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6099-6108

Automatic detection of facial Action Units (AUs) allows for objective facial expression analysis. Due to the high cost of AU labeling and the limited size of existing benchmarks, previous AU detection methods tend to overfit the dataset, resulting in a significant performance loss when evaluated across corpora. To address this problem, we propose FG-Net for generalizable facial action unit detection. Specifically, FG-Net extracts feature maps from a StyleGAN2 model pre-trained on a large and diverse face image dataset. Then, these features are used to detect AUs with a Pyramid CNN Interpreter, making the training efficient and capturing essential local features. The proposed FG-Net achieves a strong generalization ability for heatmap-based AU detection thanks to the generalizable and semantic-rich features extracted from the pre-trained generative model. Extensive experiments are conducted to evaluate within- and cross-corpus AU detection with the widely-used DISFA and BP4D datasets. Compared with the state-of-the-art, the proposed method achieves superior cross-domain performance while maintaining competitive within-domain performance. In addition, FG-Net is data-efficient and achieves competitive performance even when trained on 1000 samples. Our code will be released at https://github.com/ihp-lab/FG-Net

****

Complex Organ Mask Guided Radiology Report Generation

Tiancheng Gu, Dongnan Liu, Zhiyuan Li, Weidong Cai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7995-8004

The goal of automatic report generation is to generate a clinically accurate and coherent phrase from a single given X-ray image, which could alleviate the workload of traditional radiology reporting.However, in a real-world scenario, radiologists frequently face the challenge of producing extensive reports derived from numerous medical images, thereby medical report generation from multi-image perspective is needed.In this paper, we propose the Complex Organ Mask Guided (termed as COMG) report generation model, which incorporates masks from multiple organs (e.g., bones, lungs, heart, and mediastinum), to provide more detailed information and guide the model's attention to these crucial body regions. Specifically, we leverage prior knowledge of the disease corresponding to each organ in the fusion process to enhance the disease identification phase during the report generation process. Additionally, cosine similarity loss is introduced as target function to ensure the convergence of cross-modal consistency and facilitate model optimization.Experimental results on two public datasets show that COMG achieves a 11.4% and 9.7% improvement in terms of BLEU@4 scores over the SOTA model KiUT on IU-Xray and MIMIC, respectively.

****

Link Prediction for Flow-Driven Spatial Networks

Bastian Wittmann, Johannes C. Paetzold, Chinmay Prabhakar, Daniel Rueckert, Bjoern Menze; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2472-2481

Link prediction algorithms aim to infer the existence of connections (or links) between nodes in network-structured data and are typically applied to refine the connectivity among nodes. In this work, we focus on link prediction for flow-driven spatial networks, which are embedded in a Euclidean space and relate to physical exchange and transportation processes (e.g., blood flow in vessels or traffic flow in road networks). To this end, we propose the Graph Attentive Vectors (GAV) link prediction framework. GAV models simplified dynamics of physical flow in spatial networks via an attentive, neighborhood-aware message-passing paradigm, updating vector embeddings in a constrained manner. We evaluate GAV on eight flow-driven spatial networks given by whole-brain vessel graphs and road networks. GAV demonstrates superior performances across all datasets and metrics and outperformed the state-of-the-art on the ogbl-vessel benchmark at the time of submission by 12% (98.38 vs. 87.98 AUC). All code is publicly available on GitHub.

****

Training-Free Object Counting With Prompts
Zenglin Shi, Ying Sun, Mengmi Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 323-331

This paper tackles the problem of object counting in images. Existing approaches rely on extensive training data with point annotations for each object, making data collection labor-intensive and time-consuming. To overcome this, we propose a training-free object counter that treats the counting task as a segmentation problem. Our approach leverages the Segment Anything Model (SAM), known for its high-quality masks and zero-shot segmentation capability. However, the vanilla mask generation method of SAM lacks class-specific information in the masks, resulting in inferior counting accuracy. To overcome this limitation, we introduce a prior-guided mask generation method that incorporates three types of priors into the segmentation process, enhancing efficiency and accuracy. Additionally, we tackle the issue of counting objects specified through text by proposing a two-stage approach that combines reference object selection and prior-guided mask generation. Extensive experiments on standard datasets demonstrate the competitive performance of our training-free counter compared to learning-based approaches. This paper presents a promising solution for counting objects in various scenarios without the need for extensive data collection and counting-specific training. Code is available at https://github.com/shizenglin/training-free-object-counter.

****

SEMA: Semantic Attention for Capturing Long-Range Dependencies in Egocentric Lifelogs
Pravin Nagar, K.N. Ajay Shastry, Jayesh Chaudhari, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7025-7035

Transformer architecture is a de-facto standard for modeling global dependency in long sequences. However, quadratic space and time complexity for self-attention prohibits transformers from scaling to extremely long sequences (> 10k). Low-rank decomposition as a non-negative matrix factorization (NMF) of self-attention demonstrates remarkable performance in linear space and time complexity with strong theoretical guarantees. However, our analysis reveals that NMF-based works struggle to capture the rich spatio-temporal visual cues scattered across the long sequences resulting from egocentric lifelogs. To capture such cues, we propose a novel attention mechanism named SEMantic Attention (SEMA), which factorizes the self-attention matrix into a semantically meaningful subspace. We demonstrate SEMA in a representation learning setting, aiming to recover activity patterns in extremely long (weeks-long) egocentric lifelogs using a novel self-supervised training pipeline. Compared to the current state-of-the-art, we report significant improvement in terms of (NMI, AMI, and F-Score) for EgoRoutine, UTE, and Epic Kitchens datasets. Furthermore, to underscore the efficacy of SEMA, we extend its application to conventional video tasks such as online action detection, video recognition, and action localization.

****

Neural Image Compression Using Masked Sparse Visual Representation
Wei Jiang, Wei Wang, Yue Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4189-4197

We study neural image compression based on the Sparse Visual Representation (SVR), where images are embedded into a discrete latent space spanned by learned visual codebooks. By sharing codebooks with the decoder, the encoder transfers integer codeword indices that are efficient and cross-platform robust, and the decoder retrieves the embedded latent feature using the indices for reconstruction. Previous SVR-based compression lacks effective mechanism for rate-distortion tradeoffs, where one can only pursue either high reconstruction quality or low transmission bitrate. We propose a Masked Adaptive Codebook learning (M-AdaCode) method that applies masks to the latent feature subspace to balance bitrate and reconstruction quality. A set of semantic-class-dependent basis codebooks are learned, which are weighted combined to generate a rich latent feature for high-quality reconstruction. The combining weights are adaptively derived from each input image, providing fidelity information with additional transmission costs. By masking out unimportant weights in the encoder and recovering them in the decoder, we can trade off reconstruction quality for transmission bits, and the masking rate controls the balance between bitrate and distortion. Experiments over the standard JPEG-AI dataset demonstrate the effectiveness of our M-AdaCode approach.
****

Letting 3D Guide the Way: 3D Guided 2D Few-Shot Image Classification
Jiajing Chen, Minmin Yang, Senem Velipasalar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2732-2740
Existing few-shot image classification networks aim to perform prediction on images belonging to classes that were not seen during training, with only a few labeled images, which are randomly picked from the same image pool as the support set. However, this traditional approach has two main issues: (i) in real-world applications, since support images are randomly picked, the angle they were captured from can be very different from that of the query image, causing the images to look very different and making it hard to match them; (ii) since support and query images, for both training and testing, are sampled from the same image pool, models can overfit the dataset, especially if the image pool contains images with similar color, texture or view angle. Thus, good performance on a dataset does not reflect a model's real ability. To address these issues, we propose a novel few-shot learning approach referred to as the 3D guided 2D (3DG2D) few-shot image classification. In our proposed approach, the queries are 2D images, and the support set is composed of 3D mesh data, providing different views of an object, in contrast to randomly picked images providing a single view. From each 3D mesh, 14 projection images are generated from different angles. Thus, these projections have significant variance among themselves. To address this challenge, we also propose the Angle Inference Module (AIM), which is used to infer the view angle of a query image so that more attention is given to projection images corresponding to the same view angle as the query image to achieve better prediction performance. We perform experiments on ModelNet40, Toys4K and ShapeNet datasets with 4-fold cross validation, and show that our 3DG2D few-shot classification approach consistently outperforms the state-of-the-art baselines.
****

Specular Object Reconstruction Behind Frosted Glass by Differentiable Rendering
Takafumi Iwaguchi, Hiroyuki Kubo, Hiroshi Kawasaki; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4047-4056
This paper addresses the problem of reconstructing scenes behind optical diffusers, which is common in applications such as imaging through frosted glass. We propose a new approach that exploits specular reflection to capture sharp light distributions with a point light source, which can be used to detect reflections in low signal-to-noise scenarios. In this paper, we propose a rasterizer-based differentiable renderer to solve this problem by minimizing the difference between the captured and rendered images. Because our method can simultaneously optimize multiple observations for different light source positions, it is confirmed that ambiguities of the scene are efficiently eliminated by increasing the number of observations. Experiments show that the proposed method can reconstruct a scene with several mirror-like objects behind the diffuser in both simulated and real environments.

****

Debiasing, Calibrating, and Improving Semi-Supervised Learning Performance via Simple Ensemble Projector

Khanh-Binh Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2441-2451

Recent studies on semi-supervised learning (SSL) have achieved great success. Despite their promising performance, current state-of-the-art methods tend toward increasingly complex designs at the cost of introducing more network components and additional training procedures. In this paper, we propose a simple method named Ensemble Projectors Aided for Semi-supervised Learning (EPASS), which focuses mainly on improving the learned embeddings to boost the performance of the existing contrastive joint-training semi-supervised learning frameworks. Unlike standard methods, where the learned embeddings from one projector are stored in memory banks to be used with contrastive learning, EPASS stores the ensemble embeddings from multiple projectors in memory banks. As a result, EPASS improves generalization, strengthens feature representation, and boosts performance. For instance, EPASS improves strong baselines for semi-supervised learning by 39.47%/31.39%/24.70% top-1 error rate, while using only 100k/1%/10% of labeled data for SimMatch, and achieves 40.24%/32.64%/25.90% top-1 error rate for CoMatch on the ImageNet dataset. These improvements are consistent across methods, network architectures, and datasets, proving the general effectiveness of the proposed methods.

****

Privacy-Enhancing Person Re-Identification Framework - A Dual-Stage Approach

Kajal Kansal, Yongkang Wong, Mohan Kankanhalli; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8543-8552

In this work, we show that deep learning-based re-identification (Re-ID) models, albeit trained only with a Re-ID objective (i.e. if two samples belong to the same identity), encode personally identifiable information (PII) in the learned features that may lead to serious privacy concerns. In cognizance of the modern privacy regulations on protecting PII, we propose a novel dual-stage person Re-ID framework that (1) suppresses the PII from the discriminative features, and (2) introduces a controllable privacy mechanism through differential privacy. The former is achieved with a self-supervised de-identification (De-ID) decoder and an adversarial-identity (Adv-ID) module, whereas the latter mechanism leverages a controllable privacy budget to generate a privacy-protected gallery with a Gaussian noise generator. Furthermore, we introduce the notion of a privacy metric to quantify the privacy leakage in Re-ID features which is not explicitly examined in prior work. We demonstrate the feasibility of our approach in achieving a better trade-off between utility and privacy through rigorous experiments on person Re-ID benchmarks.

****

Detection Defenses: An Empty Promise Against Adversarial Patch Attacks on Optical Flow

Erik Scheurer, Jenny Schmalfuss, Alexander Lis, Andrés Bruhn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6489-6498

Adversarial patches undermine the reliability of optical flow predictions when placed in arbitrary scene locations. Therefore, they pose a realistic threat to real-world motion detection and its downstream applications. Potential remedies are defense strategies that detect and remove adversarial patches, but their influence on the underlying motion prediction has not been investigated. In this paper, we thoroughly examine the currently available detect-and-remove defenses ILP and LGS for a wide selection of state-of-the-art optical flow methods, and illuminate their side effects on the quality and robustness of the final flow predictions. In particular, we implement defense-aware attacks to investigate whether current defenses are able to withstand attacks that take the defense mechanism into account. Our experiments yield two surprising results: Detect-and-remove defenses do not only lower the optical flow quality on benign scenes, in doing so, they also harm the robustness under patch attacks for all tested optical flow methods except FlowNetC. As currently employed detect-and-remove defenses fail to deliver the promised adversarial robustness for optical flow, they evoke a false sense of security. The code is available at https://github.com/cv-stuttgart/DetectionDefenses.

****

SAM Fewshot Finetuning for Anatomical Segmentation in Medical Images

Weiyi Xie, Nathalie Willems, Shubham Patil, Yang Li, Mayank Kumar; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3253-3261

We propose a straightforward yet highly effective few-shot fine-tuning strategy for adapting the Segment Anything (SAM) to anatomical segmentation tasks in medical images. Our novel approach revolves around reformulating the mask decoder within SAM, leveraging few-shot embeddings derived from a limited set of labeled images (few-shot collection) as prompts for querying anatomical objects captured in image embeddings. This innovative reformulation greatly reduces the need for time-consuming online user interactions for labeling volumetric images, such as exhaustively marking points and bounding boxes to provide prompts slice by slice. With our method, users can manually segment a few 2D slices offline, and the embeddings of these annotated image regions serve as effective prompts for online segmentation tasks. Our method prioritizes the efficiency of the fine-tuning process by exclusively training the mask decoder through caching mechanisms while keeping the image encoder frozen. Importantly, this approach is not limited to volumetric medical images, but can generically be applied to any 2D/3D segmentation task. To thoroughly evaluate our method, we conducted extensive validation on four datasets, covering six anatomical segmentation tasks across two modalities. Furthermore, we conducted a comparative analysis of different prompting options within SAM and the fully-supervised nnU-Net. The results demonstrate the superior performance of our method compared to SAM employing only point prompts (50% improvement in IoU) and performs on-par with fully supervised methods whilst reducing the requirement of labeled data by at least an order of magnitude.

****

BirdSAT: Cross-View Contrastive Masked Autoencoders for Bird Species Classification and Mapping

Srikumar Sastry, Subash Khanal, Aayush Dhakal, Di Huang, Nathan Jacobs; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7136-7145

We propose a metadata-aware self-supervised learning (SSL) framework useful for fine-grained classification and ecological mapping of bird species around the world. Our framework unifies two SSL strategies: Contrastive Learning (CL) and Masked Image Modeling (MIM), while also enriching the embedding space with meta-information available with ground-level imagery of birds. We separately train uni-modal and cross-modal ViT on a novel cross-view global birds species dataset containing ground-level imagery, metadata (location, time), and corresponding satellite imagery. We demonstrate that our models learn fine-grained and geographically conditioned features of birds, by evaluating on two downstream tasks: fine-grained visual classification (FGVC) and cross-modal retrieval. Pre-trained models learned using our framework achieve SotA performance on FGVC of iNAT-2021 birds as well as in transfer learning setting for CUB-200-2011 and NABirds datasets. Moreover, the impressive cross-modal retrieval performance of our model enables the creation of species distribution maps across any geographic region. The dataset and source code will be released at https://github.com/TBD.

****

FELGA: Unsupervised Fragment Embedding for Fine-Grained Cross-Modal Association

Yaoxin Zhuo, Baoxin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5635-5645

Vision-and-Language Pre-trained (VLP) models have demonstrated their powerful zero-shot ability in multiple downstream tasks. Most of these models are designed to learn joint embeddings of images and their paired sentences, with both modalities considered globally. This does not lead to optimal solutions for applications where what matters more is the local-level cross-modal association, such as the situation where a user may want to retrieve images with query words that link to only small parts of the images. While a VLP model could in principle be retrained to learn a new embedding capturing such fine-grained association, expensive annotation would be needed, making it impractical for big data applications. This paper proposes a novel method named Fragment Embedding by Local and Global Alignment (FELGA), which learns fragment-level embeddings that capture fine-grained cross-modal association through utilizing visual entity proposals and semantic concept proposals in an unsupervised manner. Comprehensive experiments conducted on three VLP models and two datasets demonstrate that FELGA is not limited to specific VLP models and outperforms the original VLP features. In particular, the learned embeddings support cross-modal fragment association tasks including query-driven object discovery and description assignment.

****

Weakly-Supervised Deepfake Localization in Diffusion-Generated Images

Drago■-Constantin ■ân■aru, Elisabeta Onea■■, Dan Onea■■; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6258-6268

The remarkable generative capabilities of denoising diffusion models have raised new concerns regarding the authenticity of the images we see every day on the Internet. However, the vast majority of existing deepfake detection models are tested against previous generative approaches (e.g. GAN) and usually provide only a "fake" or "real" label per image. We believe a more informative output would be to augment the per-image label with a localization map indicating which regions of the input have been manipulated. To this end, we frame this task as a weakly-supervised localization problem and identify three main categories of methods (based on either explanations, local scores or attention), which we compare on an equal footing by using the Xception network as the common backbone architecture. We provide a careful analysis of all the main factors that parameterize the design space: choice of method, type of supervision, dataset and generator used in the creation of manipulated images; our study is enabled by constructing datasets in which only one of the components is varied. Our results show that weakly-supervised localization is attainable, with the best performing detection method (based on local scores) being less sensitive to the looser supervision than to the mismatch in terms of dataset or generator.

****

Sketch-Based Video Object Localization

Sangmin Woo, So-Yeong Jeon, Jinyoung Park, Minji Son, Sumin Lee, Changick Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8480-8489

We introduce Sketch-based Video Object Localization (SVOL), a new task aimed at localizing spatio-temporal object boxes in video queried by the input sketch. We first outline the challenges in the SVOL task and build the Sketch-Video Attention Network (SVANet) with the following design principles: (i) to consider temporal information of video and bridge the domain gap between sketch and video; (ii) to accurately identify and localize multiple objects simultaneously; (iii) to handle various styles of sketches; (iv) to be classification-free. In particular, SVANet is equipped with a Cross-modal Transformer that models the interaction between learnable object tokens, query sketch, and video through attention operations, and learns upon a per-frame set matching strategy that enables frame-wise prediction while utilizing global video context. We evaluate SVANet on a newly curated SVOL dataset. By design, SVANet successfully learns the mapping between the query sketches and video objects, achieving state-of-the-art results on the SVOL benchmark. We further confirm the effectiveness of SVANet via extensive ablation studies and visualizations. Lastly, we demonstrate its transfer capability on unseen datasets and novel categories, suggesting its high scalability in real-world applications. Codes are available at https://github.com/sangminwoo/SVOL.

****

Hyperbolic vs Euclidean Embeddings in Few-Shot Learning: Two Sides of the Same Coin

Gabriel Moreira, Manuel Marques, João Paulo Costeira, Alexander Hauptmann; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WAC

V), 2024, pp. 2082-2090

Recent research in representation learning has shown that hierarchical data lends itself to low-dimensional and highly informative representations in hyperbolic space. However, even if hyperbolic embeddings have gathered attention in image recognition, their optimization is prone to numerical hurdles. Further, it remains unclear which applications stand to benefit the most from the implicit bias imposed by hyperbolicity, when compared to traditional Euclidean features. In this paper, we focus on prototypical hyperbolic neural networks. In particular, the tendency of hyperbolic embeddings to converge to the boundary of the Poincare ball in high dimensions and the effect this has on few-shot classification. We show that the best few-shot results are attained for hyperbolic embeddings at a common hyperbolic radius. In contrast to prior benchmark results, we demonstrate that better performance can be achieved by a fixed-radius encoder equipped with the Euclidean metric, regardless of the embedding dimension.

****

Attention-Guided Prototype Mixing: Diversifying Minority Context on Imbalanced Whole Slide Images Classification Learning

Farchan Hakim Raswa, Chun-Shien Lu, Jia-Ching Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7624-7633

Real-world medical datasets often suffer from class imbalance, which can lead to degraded performance due to limited samples of the minority class. In another line of research, Transformer-based multiple instance learning (Transformer-MIL) has shown promise in addressing the pairwise correlation between instances in medical whole slide images (WSIs) with gigapixel resolution and non-uniform sizes. However, these characteristics pose challenges for state-of-the-art (SOTA) oversampling methods aiming at diversifying the minority context in imbalanced WSIs. In this paper, we propose an Attention-Guided Prototype Mixing scheme at the WSI level. We leverage Transformer-MIL training to determine the distribution of semantic instances and identify relevant instances for cutting and pasting across different WSI (bag of instances). To our knowledge, applying Transformer is often limited by memory requirements and time complexity, particularly when dealing with gigabyte-sized WSIs. We introduce the concept of prototype instances that have smaller representations while preserving the uniform size and intrinsic features of the WSI. We demonstrate that our proposed method can boost performance compared to competitive SOTA oversampling and augmentation methods at an imbalanced WSI level.

****

StyleAvatar: Stylizing Animatable Head Avatars

Juan C. Pérez, Thu Nguyen-Phuoc, Chen Cao, Artsiom Sanakoyeu, Tomas Simon, Pablo Arbeláez, Bernard Ghanem, Ali Thabet, Albert Pumarola; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8678-8687

AR/VR applications promise to provide people with a genuine feeling of mutual presence when communicating via their personalized avatars. While realistic avatars are essential in various social settings, the vast possibilities of a virtual world can also generate interest in using stylized avatars for other purposes. We introduce StyleAvatar, the first method for semantic stylization of animatable head avatars. StyleAvatar directly stylizes the avatar representation, rather than stylizing its renders. Specifically, given a model generating the avatar, StyleAvatar first disentangles geometry and texture manipulations, and then stylizes the avatar by fine-tuning a subset of the model's weights. Our method has multiple virtues, including the ability to describe styles using images or text, preserving the avatar's animatable capacity, providing control over identity preservation, and disentangling texture and geometry modifications. Experiments have shown that our approach consistently works across skin tones, challenging hair styles, extreme views, and diverse facial expressions.

****

On the Quantification of Image Reconstruction Uncertainty Without Training Data

Jiaxin Zhang, Sirui Bi, Victor Fung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2072-2081

Computational imaging plays a pivotal role in determining hidden information from sparse measurements. A robust inverse solver is crucial to fully characterize the uncertainty induced by these measurements, as it allows for the estimation of the complete posterior of unrecoverable targets. This, in turn, facilitates a probabilistic interpretation of observational data for decision-making. In this study, we propose a deep variational framework that leverages a deep generative model to learn an approximate posterior distribution to effectively quantify image reconstruction uncertainty without the need for training data. We parameterize the target posterior using a flow-based model and minimize their Kullback-Leibler (KL) divergence to achieve accurate uncertainty estimation. To bolster stability, we introduce a robust flow-based model with bi-directional regularization and enhance expressivity through gradient boosting. Additionally, we incorporate a space-filling design to achieve substantial variance reduction on both latent prior space and target posterior space. We validate our method on several benchmark tasks and two real-world applications, namely fastMRI and black hole image reconstruction. Our results indicate that our method provides reliable and high-quality image reconstruction with robust uncertainty estimation.

****

Sparse Convolutional Networks for Surface Reconstruction From Noisy Point Clouds

Tao Wang, Jing Wu, Ze Ji, Yu-Kun Lai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3212-3221

Reconstructing accurate 3D surfaces from noisy point clouds is a fundamental problem in computer vision. Among different approaches, neural implicit methods that map 3D coordinates to occupancy values benefit from the learning capabilities of deep neural networks and the flexible topology of implicit representations, and achieve promising reconstruction results. However, existing methods utilize standard (dense) 3D convolutional neural networks for feature extraction and occupancy prediction, which significantly restricts the capability to reconstruct details. In this paper, we propose a neural implicit method based on sparse convolutions, where features and network calculations only focus on grid points close to the surface to be reconstructed. This allows us to build significantly higher resolution 3D grids and reconstruct high-fidelity details. We further build a 3D residual UNet to extract features which are robust to noise, while ensuring details are retained. A 3D position along with features extracted at the position are fed into the occupancy probability predictor network to obtain occupancy. As features at nearby grid points to the query position may not exist due to the sparse nature, we propose a normalized weight interpolation approach to obtain smooth interpolation with sparse data. Experimental results demonstrate that our method achieves promising results, both qualitatively and quantitatively, outperforming existing methods.

****

Self-Sampling Meta SAM: Enhancing Few-Shot Medical Image Segmentation With Meta-Learning

Tianang Leng, Yiming Zhang, Kun Han, Xiaohui Xie; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7925-7935

While the Segment Anything Model (SAM) excels in semantic segmentation for general-purpose images, its performance significantly deteriorates when applied to medical images, primarily attributable to insufficient representation of medical images in its training dataset. Nonetheless, gathering comprehensive datasets and training models that are universally applicable is particularly challenging due to the long-tail problem common in medical images. To address this gap, here we present a Self-Sampling Meta SAM (SSM-SAM) framework for few-shot medical image segmentation. Our innovation lies in the design of three key modules: 1) An online fast gradient descent optimizer, further optimized by a meta-learner, which ensures swift and robust adaptation to new tasks. 2) A Self-Sampling module designed to provide well-aligned visual prompts for improved attention allocation; and 3) A robust attention-based decoder specifically designed for medical few-shot learning to capture relationship between different slices. Extensive experiments on a popular abdominal CT dataset and an MRI dataset demonstrate that the proposed method achieves significant improvements over state-of-the-art methods in few-shot segmentation, with an average improvements of 10.21% and 1.80% in terms of DSC, respectively. In conclusion, we present a novel approach for rapid online adaptation in interactive image segmentation, adapting to a new organ in just 0.83 minutes. Code is available at https://github.com/DragonDescentZerotsu/SSM-SAM

****

LIVENet: A Novel Network for Real-World Low-Light Image Denoising and Enhancement

Dhruv Makwana, Gayatri Deshmukh, Onkar Susladkar, Sparsh Mittal, Sai Chandra Teja R.; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5856-5865

Low-light image enhancement (LLIE) is the process of improving the quality of images taken in low-light conditions while striking a balance between enhancing image illumination and maintaining their natural appearance. This involves reducing noise, enhancing details, and correcting colors, all while avoiding artifacts such as halo effects or color distortions. We propose LIVENet, a novel deep neural network that jointly performs noise reduction on lowlight images and enhances illumination and texture details. LIVENet has two stages: the image enhancement stage and the refinement stage. For the image enhancement stage, we propose a Latent Subspace Denoising Block (LSDB) that uses a low-rank representation of low light features to suppress the noise and predict a noise-free grayscale image. We propose enhancing an RGB image by eliminating noise. This is done by converting it into YCbCr color space and replacing the noisy luminance (Y) channel with the predicted noise-free grayscale image. LIVENet also predicts the transmission map and atmospheric light in the image enhancement stage. LIVENet produces an enhanced image with rich color and illumination by feeding them to an atmospheric scattering model. In the refinement stage, the texture information from the grayscale image is incorporated into the improved image using a Spatial Feature Transform (SFT) layer. Experiments on different datasets demonstrate that LIVENet's enhanced images consistently outperform previous techniques across various quality metrics. The source code can be obtained from https://github.com/CandleLabAI/LiveNet.

****

Holistic Representation Learning for Multitask Trajectory Anomaly Detection

Alexandros Stergiou, Brent De Weerdt, Nikos Deligiannis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6729-6739

Video anomaly detection deals with the recognition of abnormal events in videos. Apart from the visual signal, video anomaly detection has also been addressed with skeleton sequences. We propose a holistic representation of skeleton trajectories to learn expected motions across segments at different times. Our approach uses multitask learning to reconstruct any continuous unobserved temporal segment of the trajectory allowing the extrapolation of past and future segments and the interpolation of in-between segments. We use an end-to-end attention-based encoder-decoder to encode temporally occluded trajectories, jointly learn latent representations of the occluded trajectory segments, and reconstruct trajectories of expected motions across different temporal segments. Extensive experiments over three trajectory-based video anomaly detection datasets show the advantages and effectiveness of our method with state-of-the-art results on the detection of anomalies in skeleton trajectories

****

Text-Guided Face Recognition Using Multi-Granularity Cross-Modal Contrastive Learning

Md Mahedi Hasan, Shoaib Meraj Sami, Nasser Nasrabadi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5784-5793

State-of-the-art face recognition (FR) models often experience a significant performance drop when dealing with facial images in surveillance scenarios where images are in low quality and often corrupted with noise. Leveraging facial characteristics, such as freckles, scars, gender, and ethnicity, becomes highly benefi cial in improving FR performance in such scenarios. In this paper, we introduce text-guided face recognition (TGFR) to analyze the impact of integrating facial attributes in the form of natural language descriptions. We hypothesize that adding semantic information into the loop can significantly improve the image understanding capability of an FR algorithm compared to other soft biometrics. However, learning a discriminative joint embedding within the multimodal space poses a considerable challenge due to the semantic gap in the unaligned image-text representations, along with the complexities arising from ambiguous and incoherent textual descriptions of the face. To address these challenges, we introduce a face-caption alignment module (FCAM), which incorporates cross-modal contrastive losses across multiple granularities to maximize the mutual information between local and global features of the face-caption pair. Within FCAM, we refine both facial and textual features for learning aligned and discriminative features. We also design a face-caption fusion module (FCFM) that applies fine-grained interactions and coarse-grained associations among cross-modal features. Through extensive experiments conducted on three face-caption datasets, proposed TGFR demonstrates remarkable improvements, particularly on low-quality images, over existing FR models and outperforms other related methods and benchmarks.

****

Computer Vision on the Edge: Individual Cattle Identification in Real-Time With ReadMyCow System

Moniek Smink, Haotian Liu, Dörte Döpfer, Yong Jae Lee; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7056-7065

In precision livestock farming, the individual identification of cattle is crucial to inform the decisions made to enhance animal welfare, health, and productivity. In literature, models exist that can read ear tags; however, they are not easily portable to real-world cattle production environments and make predictions mainly on still images. We propose a video-based cattle ear tag reading system, called ReadMyCow, which takes advantage of the temporal characteristics in videos to accurately detect, track, and read cattle ear tags at 25 FPS on edge devices. For each frame in a video, ReadMyCow functions in two steps. 1) Tag detection: a YOLOv5s Object Detection model and NVIDIA Deepstream Tracking Layer detect and track the tags present. 2) Tag reading: the novel WhenToRead module decides whether to read each tag, using a TRBA Scene Text Recognition model, or to use the reading from a previous frame. The system is implemented on an edge device, namely the NVIDIA Jetson AGX Orin or Xavier, making it portable to cattle production environments without external computational resources. To attain real-time speeds, ReadMyCow only reads the detected tag in the current frame if it thinks it will get a better reading when a decision metric is significantly improved in the current frame. Ideally, this means the best reading of a tag is found and stored throughout a tag's presence in the video, even when the tag becomes occluded or blurry. While testing the system at a real Midwestern dairy farm housing 9,000 cows, 96.1% of printed ear tags were accurately read by the ReadMyCow system, demonstrating its real-world commercial potential. ReadMyCow opens opportunities for informed data-driven decision-making processes on commercial cattle farms.

****

SynergyNet: Bridging the Gap Between Discrete and Continuous Representations for Precise Medical Image Segmentation

Vandan Gorade, Sparsh Mittal, Debesh Jha, Ulas Bagci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7768-7777

In recent years, continuous latent space (CLS) and discrete latent space (DLS) deep learning models have been proposed for medical image analysis for improved performance. However, these models encounter distinct challenges. CLS models capture intricate details but often lack interpretability in terms of structural representation and robustness due to their emphasis on low-level features. Conversely, DLS models offer interpretability, robustness, and the ability to capture coarse-grained information thanks to their structured latent space. However, DLS m

odels have limited efficacy in capturing fine-grained details. To address the limitations of both DLS and CLS models, we propose SynergyNet, a novel bottleneck architecture designed to enhance existing encoder-decoder segmentation frameworks. SynergyNet seamlessly integrates discrete and continuous representations to harness complementary information and successfully preserves both fine and coarse grained details in the learned representations. Our extensive experiment on multi-organ segmentation and cardiac datasets demonstrates that SynergyNet outperforms other state of the art methods including TransUNet: dice scores improving by 2.16%, and Hausdorff scores improving by 11.13%, respectively. When evaluating skin lesion and brain tumor segmentation datasets, we observe a remarkable improvements of 1.71% in Intersection-overUnion scores for skin lesion segmentation and of 8.58% for brain tumor segmentation. Our innovative approach paves the way for enhancing the overall performance and capabilities of deep learning models in the critical domain of medical image analysis.
****
Rethinking Visibility in Human Pose Estimation: Occluded Pose Reasoning via Transformers
Pengzhan Sun, Kerui Gu, Yunsong Wang, Linlin Yang, Angela Yao; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5903-5912
Occlusion is a common challenge in human pose estimation. Curiously, learning from occluded keypoints hinders a model to detect visible keypoints. We speculate that the impairment is likely due to a forced correlation between keypoints and visual features of the occluders. As such, we propose a novel visibility-aware attention mechanism to eliminate unreliable occluding features. The explicit occlusion handling encourages the model to reason about occluded keypoints using evidence and contextual information from the visible keypoints. It also mitigates the damage of unreliable correlations of the occluded keypoints. Our method, when added to the strong baseline SimCC, improves by 1.3 AP and 0.7 AP with ResNet and HRNet respectively. It also surpasses the state-of-the-art I^2R-Net on CrowdPose by 0.3 AP and 0.6 AP^hard. The improvements highlight that rethinking visibility information is critical for developing effective human pose estimation systems.
****
CCMR: High Resolution Optical Flow Estimation via Coarse-To-Fine Context-Guided Motion Reasoning
Azin Jahedi, Maximilian Luz, Marc Rivinius, Andrés Bruhn; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6899-6908
Attention-based motion aggregation concepts have recently shown their usefulness in optical flow estimation, in particular when it comes to handling occluded regions. However, due to their complexity, such concepts have been mainly restricted to coarse-resolution single-scale approaches that fail to provide the detailed outcome of high-resolution multi-scale networks. In this paper, we hence propose CCMR: a high-resolution coarse-to-fine approach that leverages attention-based motion grouping concepts to multi-scale optical flow estimation. CCMR relies on a hierarchical two-step attention-based context-motion grouping strategy that first computes global multi-scale context features and then uses them to guide the actual motion grouping. As we iterate both steps over all coarse-to-fine scales, we adapt cross covariance image transformers to allow for an efficient realization while maintaining scale-dependent properties. Experiments and ablations demonstrate that our efforts of combining multi-scale and attention-based concepts pay off. By providing highly detailed flow fields with strong improvements in both occluded and non-occluded regions, our CCMR approach not only outperforms both the corresponding single-scale attention-based and multi-scale attention-free baselines by up to 23.0% and 21.6%, respectively, it also achieves state-of-the-art results, ranking first on KITTI 2015 and second on MPI Sintel Clean and Final. Code and trained models are available at https://github.com/cv-stuttgart/CCMR.
****

ReConPatch: Contrastive Patch Representation Learning for Industrial Anomaly Detection

Jeeho Hyun, Sangyun Kim, Giyoung Jeon, Seung Hwan Kim, Kyunghoon Bae, Byung Jun Kang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2052-2061

Anomaly detection is crucial to the advanced identification of product defects such as incorrect parts, misaligned components, and damages in industrial manufacturing. Due to the rare observations and unknown types of defects, anomaly detection is considered to be challenging in machine learning. To overcome this difficulty, recent approaches utilize the common visual representations pre-trained from natural image datasets and distill the relevant features. However, existing approaches still have the discrepancy between the pre-trained feature and the target data, or require the input augmentation which should be carefully designed, particularly for the industrial dataset. In this paper, we introduce ReConPatch, which constructs discriminative features for anomaly detection by training a linear modulation of patch features extracted from the pre-trained model. ReConPatch employs contrastive representation learning to collect and distribute features in a way that produces a target-oriented and easily separable representation. To address the absence of labeled pairs for the contrastive learning, we utilize two similarity measures between data representations, pairwise and contextual similarities, as pseudo-labels. Our method achieves the state-of-the-art anomaly detection performance (99.72%) for the widely used and challenging MVTec AD dataset. Additionally, we achieved a state-of-the-art anomaly detection performance (95.8%) for the BTAD dataset.

****

Are Natural Domain Foundation Models Useful for Medical Image Classification?

Joana Palés Huix, Adithya Raju Ganeshan, Johan Fredin Haslum, Magnus Söderberg, Christos Matsoukas, Kevin Smith; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7634-7643

The deep learning field is converging towards the use of general foundation models that can be easily adapted for diverse tasks. While this paradigm shift has become common practice within the field of natural language processing, progress has been slower in computer vision. In this paper we attempt to address this issue by investigating the transferability of various state-of-the-art foundation models to medical image classification tasks. Specifically, we evaluate the performance of five foundation models, namely SAM, SEEM, DINOv2, BLIP, and OpenCLIP across four well-established medical imaging datasets. We explore different training settings to fully harness the potential of these models. Our study shows mixed results. DINOv2 consistently outperforms the standard practice of ImageNet pretraining. However, other foundation models failed to consistently beat this established baseline indicating limitations in their transferability to medical image classification tasks.

****

Favoring One Among Equals - Not a Good Idea: Many-to-One Matching for Robust Transformer Based Pedestrian Detection

K.N. Ajay Shastry, K. Ravi Sri Teja, Aditya Nigam, Chetan Arora; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 759-768

We investigate the reasons for lower performance of transformer based pedestrian detection models compared to convolutional neural network (CNN) based ones. CNN models generate dense pedestrian proposals, refine each proposal individually, and follow it up with non-maximal-suppression (NMS) to generate sparse predictions. In contrast, transformer models select one proposal per ground-truth (GT) pedestrian box and backpropagate positive gradient from them. All other proposals, many of them highly similar to the selected ones, are passed negative gradient. Though this leads to sparse predictions, obviating the need of NMS, the arbitrary selection of one among many similar proposals, hinders effective training, and lower accuracy of pedestrian detection. To mitigate the problem, instead of commonly used Kuhn-Munkres matching algorithm, we propose Min-cost-flow based formulation, and incorporate constraints such as, each ground truth box is matched t o atleast one proposal, and many equally good proposals can be matched to a single ground truth box. We propose first transformer based pedestrian detection model incorporating our matching algorithm. Extensive experiments reveal that our approach achieves a miss rate (lower is better) of 3.7 / 17.4 / 21.8 / 8.3 / 2.0 on Eurocity / TJU-traffic / TJU-campus / Cityperson / Caltech datasets compared to 4.7 / 18.7 / 24.8 / 8.5 / 3.1 by the current SOTA. Code is available at https://ajayshastry08.github.io/flow_matcher

****

MACP: Efficient Model Adaptation for Cooperative Perception

Yunsheng Ma, Juanwu Lu, Can Cui, Sicheng Zhao, Xu Cao, Wenqian Ye, Ziran Wang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3373-3382

Vehicle-to-vehicle (V2V) communications have greatly enhanced the perception capabilities of connected and automated vehicles (CAVs) by enabling information sharing to "see through the occlusions", resulting in significant performance improvements. However, developing and training complex multi-agent perception models from scratch can be expensive and unnecessary when existing single-agent models show remarkable generalization capabilities. In this paper, we propose a new framework termed MACP, which equips a single-agent pre-trained model with cooperation capabilities. We approach this objective by identifying the key challenges of shifting from single-agent to cooperative settings, adapting the model by freezing most of its parameters and adding a few lightweight modules. We demonstrate in our experiments that the proposed framework can effectively utilize cooperative observations and outperform other state-of-the-art approaches in both simulated and real-world cooperative perception benchmarks while requiring substantially fewer tunable parameters with reduced communication costs. Our source code is available at https://github.com/PurdueDigitalTwin/MACP.

****

Adapt Your Teacher: Improving Knowledge Distillation for Exemplar-Free Continual Learning

Filip Szatkowski, Mateusz Pyla, Marcin Przewi■■likowski, Sebastian Cygert, Bart■omiej Twardowski, Tomasz Trzci■ski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1977-1987

In this work, we investigate exemplar-free class incremental learning (CIL) with knowledge distillation (KD) as a regularization strategy, aiming to prevent forgetting. KD-based methods are successfully used in CIL, but they often struggle to regularize the model without access to exemplars of the training data from previous tasks. Our analysis reveals that this issue originates from substantial representation shifts in the teacher network when dealing with out-of-distribution data. This causes large errors in the KD loss component, leading to performance degradation in CIL models. Inspired by recent test-time adaptation methods, we introduce Teacher Adaptation (TA), a method that concurrently updates the teacher and the main models during incremental training. Our method seamlessly integrates with KD-based CIL approaches and allows for consistent enhancement of their performance across multiple exemplar-free CIL benchmarks. The source code for our method is available at https://github.com/fszatkowski/cl-teacher-adaptation.

****

Content-Aware Image Color Editing With Auxiliary Color Restoration Tasks

Yixuan Ren, Jing Shi, Zhifei Zhang, Yifei Fan, Zhe Lin, Bo He, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5192-5201

Diversified image color editing is typically modeled as a multimodal image-to-image translation (MMI2IT) problem with an impact on multiple applications such as photo enhancement and retouching. Although previous GAN-based algorithms successfully generate diverse edits with clear control, we observe two issues remaining: Firstly, they tend to apply the same color style to all kinds of input images when sampling with the same style latent, regardless of the input content and scenes. Secondly, they usually edit the color style globally in an image and fail to keep each semantic region and instance in harmonic colors individually. We attribute these issues to the strong independence between the style latent and th e condition image in most current MMI2IT methods. To edit the raw image into a more harmonic direction with awareness of its global content and local semantics, we introduce auxiliary color restoration tasks by reducing the input color information and training jointly. We also increase the model's capacity and enrich the noise's locality with diffusion models. Furthermore, we propose a new set of metrics to measure the content-awareness of MMI2IT models, that is, how the generated style is adaptive to the input image's content. Our model is also extensible to several downstream applications including exemplar-based color editing and language-guided color editing, without imposing extra demands on the already trained model.

****

Self-Supervised Representation Learning With Cross-Context Learning Between Global and Hypercolumn Features

Zheng Gao, Chen Feng, Ioannis Patras; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1773-1783

Whilst contrastive learning yields powerful representations by matching different augmented views of the same instance, it lacks the ability to capture the similarities between different instances. One popular way to address this limitation is by learning global features (after the global pooling) to capture inter-instance relationships based on knowledge distillation, where the global features of the teacher are used to guide the learning of the global features of the student. Inspired by cross-modality learning, we extend this existing framework that only learns from global features by encouraging the global features and intermediate layer features to learn from each other. This leads to our novel self-supervised framework: cross-context learning between global and hypercolumn features (CGH), that enforces the consistency of instance relations between low- and high-level semantics. Specifically, we stack the intermediate feature maps to construct a "hypercolumn" representation so that we can measure instance relations using two contexts (hypercolumn and global feature) separately, and then use the relations of one context to guide the learning of the other. This cross-context learning allows the model to learn from the differences between the two contexts. The experimental results on linear classification and downstream tasks show that our method outperforms the state-of-the-art methods.

****

Constrained Probabilistic Mask Learning for Task-Specific Undersampled MRI Reconstruction

Tobias Weber, Michael Ingrisch, Bernd Bischl, David Rügamer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7665-7674

Undersampling is a common method in Magnetic Resonance Imaging (MRI) to subsample the number of data points in k-space, reducing acquisition times at the cost of decreased image quality. A popular approach is to employ undersampling patterns following various strategies, e.g., variable density sampling or radial trajectories. In this work, we propose a method that directly learns the undersampling masks from data points, thereby also providing task- and domain-specific patterns. To solve the resulting discrete optimization problem, we propose a general optimization routine called ProM: A fully probabilistic, differentiable, versatile, and model-free framework for mask optimization that enforces acceleration factors through a convex constraint. Analyzing knee, brain, and cardiac MRI datasets with our method, we discover that different anatomic regions reveal distinct optimal undersampling masks, demonstrating the benefits of using custom masks, tailored for a downstream task. For example, ProM can create undersampling masks that maximize performance in downstream tasks like segmentation with networks trained on fully-sampled MRIs. Even with extreme acceleration factors, ProM yields reasonable performance while being more versatile than existing methods, paving the way for data-driven all-purpose mask generation.

****

CPSeg: Finer-Grained Image Semantic Segmentation via Chain-of-Thought Language Prompting

Lei Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Compute r Vision (WACV), 2024, pp. 513-522

Natural scene analysis and remote sensing imagery offer immense potential for advancements in large-scale language-guided context-aware data utilization. This potential is particularly significant for enhancing performance in downstream tasks such as object detection and segmentation with designed language prompting. In light of this, we introduce the CPSeg (Chain-of-Thought Language Prompting for Finer-grained Semantic Segmentation), an innovative framework designed to augment image segmentation performance by integrating a novel "Chain-of-Thought" process that harnesses textual information associated with images. This groundbreaking approach has been applied to a flood disaster scenario. CPSeg encodes prompt texts derived from various sentences to formulate a coherent chain-of-thought. We use a new vision-language dataset, FloodPrompt, which includes images, semantic masks, and corresponding text information. This not only strengthens the semantic understanding of the scenario but also aids in the key task of semantic segmentation through an interplay of pixel and text matching maps. Our qualitative and quantitative analyses validate the effectiveness of CPSeg.

****

Hyb-NeRF: A Multiresolution Hybrid Encoding for Neural Radiance Fields

Yifan Wang, Yi Gong, Yuan Zeng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3689-3698

Recent advances in Neural radiance fields (NeRF) have enabled high-fidelity scene reconstruction for novel view synthesis. However, NeRF requires hundreds of network evaluations per pixel to approximate a volume rendering integral, making it slow to train. Caching NeRFs into explicit data structures can effectively enhance rendering speed but at the cost of higher memory usage. To address these issues, we present Hyb-NeRF, a novel neural radiance field with a multi-resolution hybrid encoding that achieves efficient neural modeling and fast rendering, which also allows for high-quality novel view synthesis. The key idea of Hyb-NeRF is to represent the scene using different encoding strategies from coarse-to-fine resolution levels. Hyb-NeRF exploits coherence and compact memory of learnable positional features at coarse resolutions and the fast optimization speed and local details of hash-based feature grids at fine resolutions. In addition, to further boost performance, we embed cone tracing-based Fourier features in our learnable positional encoding that eliminates encoding ambiguity and reduces aliasing artifacts. Extensive experiments on both synthetic and real-world datasets show that Hyb-NeRF achieves faster rendering speed with better rending quality and even a lower memory footprint in comparison to previous state-of-the-art methods.

****

SequenceMatch: Revisiting the Design of Weak-Strong Augmentations for Semi-Supervised Learning

Khanh-Binh Nguyen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 96-106

Semi-supervised learning (SSL) has become popular in recent years because it allows the training of a model using a large amount of unlabeled data. However, one issue that many SSL methods face is the confirmation bias, which occurs when the model is overfitted to the small labeled training dataset and produces overconfident, incorrect predictions. To address this issue, we propose SequenceMatch, an efficient SSL method that utilizes multiple data augmentations. The key element of SequenceMatch is the inclusion of a medium augmentation for unlabeled data. By taking advantage of different augmentations and the consistency constraints between each pair of augmented examples, SequenceMatch helps reduce the divergence between the prediction distribution of the model for weakly and strongly augmented examples. In addition, SequenceMatch defines two different consistency constraints for high and low-confidence predictions. As a result, SequenceMatch is more data-efficient than ReMixMatch, and more time-efficient than both ReMixMatch (x4) and CoMatch (x2) while having higher accuracy. Despite its simplicity, SequenceMatch consistently outperforms prior methods on standard benchmarks, such as CIFAR-10/100, SVHN, and STL-10. It also surpasses prior state-of-the-art methods by a large margin on large-scale datasets such as ImageNet, with a 38.46% e rror rate.

****

Robust Learning via Conditional Prevalence Adjustment

Minh Nguyen, Alan Q. Wang, Heejong Kim, Mert R. Sabuncu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2741-2750

Healthcare data often come from multiple sites in which the correlations between confounding variables can vary widely. If deep learning models exploit these unstable correlations, they might fail catastrophically in unseen sites. Although many methods have been proposed to tackle unstable correlations, each has its limitations. For example, adversarial training forces models to completely ignore unstable correlations, but doing so may lead to poor predictive performance. Other methods (e.g. Invariant Risk Minimization) try to learn domain-invariant representations that rely only on stable associations by assuming a causal data-generating process (input X causes class label Y). Thus, they may be ineffective for anti-causal tasks (Y causes X), which are common in computer vision. We propose a method called CoPA (Conditional Prevalence-Adjustment) for anti-causal tasks. CoPA assumes that (1) generation mechanism is stable, i.e. label Y and confounding variable(s) Z generate X, and (2) the unstable conditional prevalence in each site E fully accounts for the unstable correlations between X and Y. Our crucial observation is that confounding variables are routinely recorded in healthcare settings and the prevalence can be readily estimated, for example, from a set of (Y, Z) samples (no need for corresponding samples of X). CoPA can work even if there is a single training site, a scenario which is often overlooked by existing methods. Our experiments on synthetic and real data show CoPA beating competitive baselines.

****

GRIT: GAN Residuals for Paired Image-to-Image Translation

Saksham Suri, Moustafa Meshry, Larry S. Davis, Abhinav Shrivastava; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4965-4975

Current Image-to-Image translation (I2I) frameworks rely heavily on reconstruction losses, where the output needs to match a given ground truth image. An adversarial loss is commonly utilized as a secondary loss term, mainly to add more realism to the output. Compared to unconditional GANs, I2I translation frameworks have more supervisory signals, but still their output shows more artifacts and does not reach the same level of realism achieved by unconditional GANs. We study the performance gap, in terms of photo-realism, between I2I translation and unconditional GAN frameworks. Based on our observations, we propose a modified architecture and training objective to address this realism gap. Our proposal relaxes the role of reconstruction losses, to act as regularizers instead of doing all the heavy lifting which is common in current I2I frameworks. Furthermore, our proposed formulation decouples the optimization of reconstruction and adversarial objectives and removes pixel-wise constraints on the final output. This allows for a set of stochastic but realistic variations of any target output image.

****

Embodied Human Activity Recognition

Sha Hu, Yu Gong, Greg Mori; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6447-6457

We study how to utilize the mobility of an embodied agent to improve its ability to recognize human activities. We introduce the embodied human activity recognition problem, where an agent moves in a 3D environment to recognize the category of ongoing human activities. The agent must make movement decisions based on its egocentric observations acquired up to the current time, with the goal of choosing movements to obtain new views that lead to accurate human activity recognition. Towards this goal, we propose a reinforcement learning approach that learns a policy controlling the agent's movements over time. We evaluate our approach with two realistic human activity datasets. Results show that our approach can learn to move effectively to achieve high performance in recognizing human activities.

****

INCODE: Implicit Neural Conditioning With Prior Knowledge Embeddings

Amirhossein Kazerouni, Reza Azad, Alireza Hosseini, Dorit Merhof, Ulas Bagci; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1298-1307

Implicit Neural Representations (INRs) have revolutionized signal representation by leveraging neural networks to provide continuous and smooth representations of complex data. However, existing INRs face limitations in capturing fine-grained details, handling noise, and adapting to diverse signal types. To address these challenges, we introduce INCODE, a novel approach that enhances the control of the sinusoidal-based activation function in INRs using deep prior knowledge. INCODE comprises a harmonizer network and a composer network, where the harmonizer network dynamically adjusts key parameters of the activation function. Through a task-specific pre-trained model, INCODE adapts the task-specific parameters to optimize the representation process. Our approach not only excels in representation, but also extends its prowess to tackle complex tasks such as audio, image, and 3D shape reconstructions, as well as intricate challenges such as neural radiance fields (NeRFs), and inverse problems, including denoising, super-resolution, inpainting, and CT reconstruction. Through comprehensive experiments, INCODE demonstrates its superiority in terms of robustness, accuracy, quality, and convergence rate, broadening the scope of signal representation.

****

Effective Restoration of Source Knowledge in Continual Test Time Adaptation

Fahim Faisal Niloy, Sk Miraj Ahmed, Dripta S. Raychaudhuri, Samet Oymak, Amit K. Roy-Chowdhury; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2091-2100

Traditional test-time adaptation (TTA) methods face significant challenges in adapting to dynamic environments characterized by continuously changing long-term target distributions. These challenges primarily stem from two factors: catastrophic forgetting of previously learned valuable source knowledge and gradual error accumulation caused by miscalibrated pseudo labels. To address these issues, this paper introduces an unsupervised domain change detection method that is capable of identifying domain shifts in dynamic environments and subsequently resets the model parameters to the original source pre-trained values. By restoring the knowledge from the source, it effectively corrects the negative consequences arising from the gradual deterioration of model parameters caused by ongoing shifts in the domain. Our method involves progressive estimation of global batch-norm statistics specific to each domain, while keeping track of changes in the statistics triggered by domain shifts. Importantly, our method is agnostic to the specific adaptation technique employed and thus, can be incorporated to existing TTA methods to enhance their performance in dynamic environments. We perform extensive experiments on benchmark datasets to demonstrate the superior performance of our method compared to state-of-the-art adaptation methods.

****

Unsupervised Model-Based Learning for Simultaneous Video Deflickering and Deblotching

Anuj Fulari, Satish Mulleti, Ajit Rajwade; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4117-4125

Vintage videos, as well as modern day videos acquired at high frame rates, suffer from a visually disturbing artifact called flicker, which is the rapid change in average intensity across consecutive frames. Vintage videos also suffer from blotch artifacts, i.e., each video frame contains small regions at random locations with undefined pixel values. We present a model-based learning approach to remove flicker as well as blotches simultaneously. Our work uses a pixel-wise affine intensity model for flicker between neighboring frames, with coefficients that vary smoothly in the spatial sense but randomly across time. Due to smooth spatial variation, the flicker coefficients for any given frame can be modelled as linear combinations of low frequency discrete cosine transform (DCT) bases. We also model blotches as heavy-tailed but sparse artifacts affecting every frame. We then present a novel framework to restore the video frames by jointly estimat ing the blotches as well as the DCT coefficients of the flicker, via convex optimization. Given the high computational cost of the optimization based method for processing an entire video, we use a deep unrolled neural network approach to achieve similar restoration quality at significantly reduced cost. Our approach is completely unsupervised and model based, and hence simple and interpretable. It produces high quality reconstructions, in terms of visual appeal as well as numerical metrics, on a variety of vintage videos as well as high speed videos. It does not suffer from generalization issues unlike some recent state of the art supervised methods which use end to end neural networks for restoration.

****

Gradual Source Domain Expansion for Unsupervised Domain Adaptation

Thomas Westfechtel, Hao-Wei Yeh, Dexuan Zhang, Tatsuya Harada; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1946-1955

Unsupervised domain adaptation (UDA) tries to overcome the need of a large labeled dataset by transferring knowledge from a source dataset, with lots of labeled data, to a target dataset, that has no labeled data. Since there are no labels in the target domain, early misalignment might propagate into the later stages and lead to an error build-up. In order to overcome this problem, we propose a gradual source domain expansion (GSDE) algorithm. GSDE trains the UDA task several times from scratch, but each time expands the source dataset with target data. In particular, the highest scoring target data of the previous run are employed as pseudo-source samples with their respective pseudo-label. Using this strategy, the pseudo source samples induce knowledge extracted from the previous run directly from the start of the new training. This helps align the two domains better especially in the early training epochs. In this study, we first introduce a strong baseline network and apply our GSDE strategy to it. We conduct experiments and ablation studies on three benchmarks (Office-31, OfficeHome, and DomainNet) and outperform state-of-the-art methods. We further show that the proposed GSDE strategy can improve the accuracy of a variety of different state-of-the-art UDA approaches.

****

Towards Better Structured Pruning Saliency by Reorganizing Convolution

Xinglong Sun, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2204-2214

We present SPSRC, a novel, simple and effective framework to extract enhanced Structured Pruning Saliency scores by Reorganizing Convolution. We observe that performance of pruning methods have gradually plateaued recently and propose to make better use of the learned convolutional kernel weights simply after a few steps of transformations. We firstly re-organize the convolutional operations between layers as matrix multiplications and then use the singular values and the matrix norms of the transformed matrices as saliency scores to decide what channels to prune or keep. We show both analytically and empirically that the long-standing kernel-norm-based channel importance measurement, like L1 magnitude, is not precise enough possessing a very obvious weakness of lacking spatial saliency but can be improved with SPSRC. We conduct extensive experiments across different settings and configurations and compare with the counterparts without our SPSRC along with other popular methods, observing obvious improvements. Our code is available at https://github.com/AlexSunNik/SPSRC/tree/main.

****

Controllable Text-to-Image Synthesis for Multi-Modality MR Images

Kyuri Kim, Yoonho Na, Sung-Joon Ye, Jimin Lee, Sung Soo Ahn, Ji Eun Park, Hwiyoung Kim; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7936-7945

Generative modeling has seen significant advancements in recent years, especially in the realm of text-to-image synthesis. Despite this progress, the medical field has yet to fully leverage the capabilities of large-scale foundational models for synthetic data generation. This paper introduces a framework for text-conditional magnetic resonance (MR) imaging generation, addressing the complexities associated with multi-modality considerations. The framework comprises a pre-tra ined large language model, a diffusion-based prompt-conditional image generation architecture, and an additional denoising network for input structural binary masks. Experimental results demonstrate that the proposed framework is capable of generating realistic, high-resolution, and high-fidelity multi-modal MR images that align with medical language text prompts. Further, the study interprets the cross-attention maps of the generated results based on text-conditional statements. The contributions of this research lay a robust foundation for future studies in text-conditional medical image generation and hold significant promise for accelerating advancements in medical imaging research.

****

CATS: Combined Activation and Temporal Suppression for Efficient Network Inference

Zeqi Zhu, Arash Pourtaherian, Luc Waeijen, Ibrahim Batuhan Akkaya, Egor Bondarev, Orlando Moreira; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8166-8175

Brain-inspired event-driven processors execute deep neural networks (DNNs) in an event sparsity-aware manner, leading to superior performance compared to conventional platforms. In the pursuit of higher event sparsity, prior studies suppress non-zero events by either eliminating the intra-frame activations (spatially) or leveraging the redundancy in the inter-frame differences for a video (temporally). However, we have empirically observed that simultaneously enhancing activation and temporal sparsity can lead to a synergistic suppression outcome. To this end, we propose an end-to-end event suppression training approach CATS -- Combined Activation and Temporal Suppression for efficient network inference. It utilizes a gradient-based method to search for the optimal temporal thresholds for the network while penalizing the presence of non-zero events in spatial and temporal domains simultaneously. We demonstrate that CATS achieves 2 6 times more event suppression compared to the inherent ReLU suppression, consistently outperforming the SOTA by a significant margin at various accuracy levels. Extensive experimental results show that CATS also generalizes to multiple tasks -- object detection, object tracking, pose estimation, and semantic segmentation. Furthermore, a case study for the commercial event-driven processor GrAI-VIP highlights that the induced event sparsity in SSD on EgoHands datasets efficiently translates into significant improvements of 2.5 x in FPS, 2.1 x in latency, and 3.8 x in energy consumption, while maintaining the model accuracy.

****

Learnable Cube-Based Video Encryption for Privacy-Preserving Action Recognition

Yuchi Ishikawa, Masayoshi Kondo, Hirokatsu Kataoka; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7003-7013

With the development of cloud services and machine learning, there has been an inevitable need to enhance privacy and security when serving video recognition models. Although existing image encryption methods can be used to address this issue, applying them frame by frame to videos is insufficient in two respects: model performance degradation and security strength. In this paper, we propose a novel encryption approach for privacy-preserving action recognition. It consists of two encrypting operations; Learnable Cube-based Video Encryption (LCVE) and ViT Scrambling. LCVE is video encryption based on spatio-temporal cubes, which has a large key space and can provide robust privacy protection. ViT Scrambling encrypts the Vision Transformer (ViT) model, which enables it to recognize the encrypted videos in the same manner as unencrypted videos without modifying the model architecture or fine-tuning on the encrypted data. We evaluate our method in an action recognition task with seven datasets containing a variety of action classes as well as motion and visual patterns. Empirical results demonstrate that LCVE combined with ViT Scrambling can preserve video privacy while recognizing action in encrypted videos as well as unencrypted videos. As a result, our approach outperforms existing privacy-preserving action recognition methods.

****

Learning To Generate Training Datasets for Robust Semantic Segmentation

Marwane Hariat, Olivier Laurent, Rémi Kazmierczak, Shihao Zhang, Andrei Bursuc, Angela Yao, Gianni Franchi; Proceedings of the IEEE/CVF Winter Conference on App lications of Computer Vision (WACV), 2024, pp. 3894-3905

Semantic segmentation methods have advanced significantly. Still, their robustness to real-world perturbations and object types not seen during training remains a challenge, particularly in safety-critical applications. We propose a novel approach to improve the robustness of semantic segmentation techniques by leveraging the synergy between label-to-image generators and image-to-label segmentation models. Specifically, we design Robusta, a novel robust conditional generative adversarial network to generate realistic and plausible perturbed images that can be used to train reliable segmentation models. We conduct in-depth studies of the proposed generative model, assess the performance and robustness of the downstream segmentation network, and demonstrate that our approach can significantly enhance the robustness in the face of real-world perturbations, distribution shifts, and out-of-distribution samples. Our results suggest that this approach could be valuable in safety-critical applications, where the reliability of perception modules such as semantic segmentation is of utmost importance and comes with a limited computational budget in inference. We release our code at https://github.com/ENSTA-U2IS/robusta.

****

Stereo Conversion With Disparity-Aware Warping, Compositing and Inpainting

Lukas Mehl, Andrés Bruhn, Markus Gross, Christopher Schroers; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4260-4269

Despite of exciting advances in image-based rendering and novel view synthesis, it is still challenging to achieve high-resolution results that can reach production-level quality when applying such methods to the task of stereo conversion. At the same time, only very few dedicated stereo conversion approaches exist, which also fall short in terms of the required quality. Hence, in this paper, we present a novel method for high-resolution 2D-to-3D conversion. It is fully differentiable in all of its stages and performs disparity-informed warping, consistent foreground-background compositing, and background-aware inpainting. To enable temporal consistency in the resulting video, we propose a strategy to integrate information from additional video frames. Extensive ablation studies validate our design choices, leading to a fully automatic model that outperforms existing approaches by a large margin (49-70% LPIPS error reduction). Finally, inspired from current practices in manual stereo conversion, we introduce optional interactive tools into our model, which allow to steer the conversion process and make it significantly more applicable for 3D film production.

****

GTP-ViT: Efficient Vision Transformers via Graph-Based Token Propagation

Xuwei Xu, Sen Wang, Yudong Chen, Yanping Zheng, Zhewei Wei, Jiajun Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 86-95

Vision Transformers (ViTs) have revolutionized the field of computer vision, yet their deployments on resource-constrained devices remain challenging due to high computational demands. To expedite pre-trained ViTs, token pruning and token merging approaches have been developed, which aim at reducing the number of tokens involved in the computation. However, these methods still have some limitations, such as image information loss from pruned tokens and inefficiency in the token-matching process. In this paper, we introduce a novel Graph-based Token Propagation (GTP) method to resolve the challenge of balancing model efficiency and information preservation for efficient ViTs. Inspired by graph summarization algorithms, GTP meticulously propagates less significant tokens' information to spatially and semantically connected tokens that are of greater importance. Consequently, the remaining few tokens serve as a summarization of the entire token graph, allowing the method to reduce computational complexity while preserving essential information of eliminated tokens. Combined with an innovative token selection strategy, GTP can efficiently identify image tokens to be propagated. Extensive experiments have validated GTP's effectiveness, demonstrating both efficiency and performance improvements. Specifically, GTP decreases the computational complexity of both DeiT-S and DeiT-B by up to 26% with only a minimal 0.3% accuracy drop on ImageNet-1K without finetuning, and remarkably surpasses the state-of-the-art token merging method on various backbones at an even faster inference speed. The source code is available in the supplementary material.

****

HaGRID -- HAnd Gesture Recognition Image Dataset
Alexander Kapitanov, Karina Kvanchiani, Alexander Nagaev, Roman Kraynov, Andrei Makhliarchuk; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4572-4581

This paper introduces an enormous dataset, HaGRID (HAnd Gesture Recognition Image Dataset), to build a hand gesture recognition (HGR) system concentrating on interaction with devices to manage them. That is why all 18 chosen gestures are endowed with the semiotic function and can be interpreted as a specific action. Although the gestures are static, they were picked up, especially for the ability to design several dynamic gestures. It allows the trained model to recognize not only static gestures such as 'like' and 'stop' but also 'swipes' and 'drag and drop' dynamic gestures. The HaGRID contains 554,800 images and bounding box annotations with gesture labels to solve hand detection and gesture classification tasks. The low variability in context and subjects of other datasets was the reason for creating the dataset without such limitations. Utilizing crowdsourcing platforms allowed us to collect samples recorded by 37,583 subjects in at least as many scenes with subject-to-camera distances from 0.5 to 4 meters in various natural light conditions. The influence of the diversity characteristics was assessed in ablation study experiments. Also, we demonstrate the HaGRID ability to be used for pretraining models in HGR tasks. The HaGRID and pre-trained models are publicly available.

****

Beyond RGB: A Real World Dataset for Multispectral Imaging in Mobile Devices
Ortal Glatt, Yotam Ater, Woo-Shik Kim, Shira Werman, Oded Berby, Yael Zini, Shay Zelinger, Sangyoon Lee, Heejin Choi, Evgeny Soloveichik; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4344-4354

Multispectral (MS) imaging systems have a wide range of applications for computer vision and computational photography tasks, but do not yet enjoy widespread adoption due to their prohibitive costs. Recently, advances in the design and fabrication of photonic metamaterials have enabled the development of MS sensors suitable for integration into consumer grade mobile devices. Augmenting existing RGB cameras and their processing algorithms with richer spectral information has the potential to be utilized in many steps of the image processing pipeline, but diverse real world datasets suitable for conducting such research are not freely available. We introduce Beyond RGB, a real-world dataset comprising thousands of multispectral and RGB images in diverse real world and lab conditions that is suitable for the development and evaluation of algorithms utilizing multispectral and RGB data. All the scenes in our dataset include a colorimetric reference and a measurement of the spectrum of the scene illuminant. Additionally, we demonstrate the practical use of our dataset through the introduction of a novel illuminant spectral estimation (ISE) algorithm. Our algorithm surpasses the current state-of-the-art (SoTA) by up to 58.6% on established benchmarks and sets a baseline performance on our own dataset.

****

Lightweight Portrait Matting via Regional Attention and Refinement
Yatao Zhong, Ilya Zharkov; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4158-4167

We present a lightweight model for high resolution portrait matting. The model does not use any auxiliary inputs such as trimaps or background captures and achieves real time performance for HD videos and near real time for 4K. Our model is built upon a two-stage framework with a low resolution network for coarse alpha estimation followed by a refinement network for local region improvement. However, a naive implementation of the two-stage model suffers from poor matting quality if not utilizing any auxiliary inputs. We address the performance gap by leveraging the vision transformer (ViT) as the backbone of the low resolution netwo rk, motivated by the observation that the tokenization step of ViT can reduce spatial resolution while retain as much pixel information as possible. To inform local regions of the context, we propose a novel cross region attention (CRA) module in the refinement network to propagate the contextual information across the neighboring regions. We demonstrate that our method achieves superior results and outperforms other baselines on three benchmark datasets while only uses 1/20 of the FLOPS compared to the existing state-of-the-art model.

****

Analyzing the Domain Shift Immunity of Deep Homography Estimation

Mingzhen Shao, Tolga Tasdizen, Sarang Joshi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4800-4808

Homography estimation serves as a fundamental technique for image alignment in a wide array of applications. The advent of convolutional neural networks has introduced learning-based methodologies that have exhibited remarkable efficacy in this realm. Yet, the generalizability of these approaches across distinct domains remains underexplored. Unlike other conventional tasks, CNN-driven homography estimation models show a distinctive immunity to domain shifts, enabling seamless deployment from one dataset to another without the necessity of transfer learning. This study explores the resilience of a variety of deep homography estimation models to domain shifts, revealing that the network architecture itself is not a contributing factor to this remarkable adaptability. By closely examining the models' focal regions and subjecting input images to a variety of modifications, we confirm that the models heavily rely on local textures such as edges and corner points for homography estimation. Moreover, our analysis underscores that the domain shift immunity itself is intricately tied to the utilization of these local textures.

****

Gradient Coreset for Federated Learning

Durga Sivasubramanian, Lokesh Nagalapatti, Rishabh Iyer, Ganesh Ramakrishnan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2648-2657

Federated Learning (FL) is used to learn machine learning models with data that is partitioned across multiple clients, including resource-constrained edge devices. It is therefore important to devise solutions that are efficient in terms of compute, communication, and energy consumption, while ensuring compliance with the FL framework's privacy requirements. Conventional approaches to these problems select a weighted subset of the training dataset, known as coreset, and learn by fitting models on it. Such coreset selection approaches are also known to be robust to data noise. However, these approaches rely on the overall statistics of the training data and are not easily extendable to the FL setup. In this paper, we propose an algorithm called Gradient based Coreset for Robust and Efficient Federated Learning (GCFL) that selects a coreset at each client, only every K communication rounds and derives updates only from it, assuming the availability of a small validation dataset at the server. We demonstrate that our coreset selection technique is highly effective in accounting for noise in clients' data. We conduct experiments using four real-world datasets and show that GCFL is (1) more compute and energy efficient than FL, (2) robust to various kinds of noise in both the feature space and labels, (3) preserves the privacy of the validation dataset, and (4) introduces a small communication overhead but achieves significant gains in performance, particularly in cases when the clients' data is noisy.

****

Semantic Fusion Augmentation and Semantic Boundary Detection: A Novel Approach to Multi-Target Video Moment Retrieval

Cheng Huang, Yi-Lun Wu, Hong-Han Shuai, Ching-Chun Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6783-6792

Given an untrimmed video and a natural language query, video moment retrieval (VMR) aims to retrieve video moments described by the query. However, most existing VMR methods assume a one-to-one mapping between the input query and the target video moment (single-target VMR), disregarding the possibility that a video may contain multiple target moments that match the query description (multi-target VMR). Previous methods tackle multi-target VMR by incorporating false negative moments with the original target moment for multi-target training. However, existing methods cannot properly work when no false negative moments exist in the video, or when the identified false negative moments are noisy but are still being utilized as pseudo-labels. In this paper, we propose to tackle multi-target VMR by Semantic Fusion Augmentation and Semantic Boundary Detection (SFABD). Specifically, we use feature-level augmentation to generate augmented target moments, along with an intra-video contrastive loss to ensure feature consistency. Meanwhile, we perform semantic boundary detection to adaptively remove all false negatives from the negative set of contrastive loss to avoid semantic confusion. Extensive experiments conducted on Charades-STA, ActivityNet Captions, and QVHighlights show that our method achieves state-of-the-art performance on multi-target metrics and single-target metrics. The source code is available at https://github.com/basiclab/SFABD.

****

CHAI: Craters in Historical Aerial Images

Marvin Burges, Sebastian Zambanini, Philipp Pirker; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8256-8265

In this paper we highlight the importance of historical aerial images in better understanding past events and their impact on their surroundings. More specifically, we are interested in studying bomb craters from World War II in Central Europe. We note the scarcity of publicly accessible datasets that provide labeled bomb craters and subsequently introduce a novel, domain-expert-annotated dataset comprised of 99 historical aerial images of Austria and Germany. We divide said data into training, validation, and test sets, and conduct training and evaluation using different object detectors - both general purpose and specifically designed for remote sensing applications. This dataset thus serves as a benchmark for developing and evaluating (several) algorithms dedicated to the automated detection and analysis of bomb craters in historical aerial images. We underscore the uniqueness of this dataset as the first publicly available resource containing annotated bomb craters, thereby offering researchers a valueable and novel opportunity for future exploration. Lastly, we investigate possibilities for extending and enriching our data to enhance future studies, particularly within the context of preliminary risk estimation for unexploded bombs.

****

PDA-RWSR: Pixel-Wise Degradation Adaptive Real-World Super-Resolution

Andreas Aakerberg, Majed El Helou, Kamal Nasrollahi, Thomas Moeslund; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4097-4107

While many methods have been proposed to solve the Super-Resolution (SR) problem of Low-Resolution (LR) images with complex unknown degradations, their performance still drops significantly when evaluated on images with challenging real-world degradations. One often overlooked factor contributing to this, is the presence of spatially varying degradations in real LR images. To address this issue, we propose a novel degradation pipeline capable of generating paired LR/High-Resolution (HR) images with spatially varying noise, a key contributor to reducedimage quality. Furthermore, to fully leverage such training data, we novelly propose a Pixel-Wise Degradation Adaptive Real-World Super-Resolution (PDA-RWSR) framework. Specifically, we design a new Restormer-based Real-World Super-Resolution (RWSR) model capable of adapting the reconstruction process based on pixel-wise degradation features extracted by a new supervised degradation estimation model. Along with our proposed method, we also introduce a new challenging real-world Spatially Variant Super-Resolution (SVSR) benchmarking dataset, where the images are degraded by complex noise of varying intensity and type, to evaluate the robustness of existing RWSR methods. Comprehensive experiments on synthetic and the proposed challenging real dataset demonstrates the superiority of our method over the current State-of-The-Art (SoTA).

****

Zero-Shot Video Moment Retrieval From Frozen Vision-Language Models

Dezhao Luo, Jiabo Huang, Shaogang Gong, Hailin Jin, Yang Liu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5464-5473

Accurate video moment retrieval (VMR) requires universal visual-textual correlations that can handle unknown vocabulary and unseen scenes. However, the learned correlations are likely either biased when derived from a limited amount of moment-text data which is hard to scale up because of the prohibitive annotation cost (fully-supervised), or unreliable when only the video-text pairwise relationships are available without fine-grained temporal annotations (weakly supervised). Recently, the vision-language models (VLM) demonstrate a new transfer learning paradigm to benefit different vision tasks through the universal visual-textual correlations derived from large-scale vision-language pairwise web data, which has also shown benefits to VMR by fine-tuning in the target domains. In this work, we propose a zero-shot method for adapting generalisable visual textual priors from arbitrary VLM to facilitate moment-text alignment, without the need for accessing the VMR data. To this end, we devise a conditional feature refinement module to generate boundary-aware visual features conditioned on text queries to enable better moment boundary understanding. Additionally, we design a bottom-up proposal generation strategy that mitigates the impact of domain discrepancies and breaks down complex-query retrieval tasks into individual action retrievals, thereby maximizing the benefits of VLM. Extensive experiments conducted on three VMR benchmark datasets demonstrate the notable performance advantages of our zero-shot algorithm, especially in the novel-word and novel-location out-of-distribution setups.

****

Image Denoising and the Generative Accumulation of Photons

Alexander Krull, Hector Basevi, Benjamin Salmon, Andre Zeug, Franziska Müller, Samuel Tonks, Leela Muppala, Aleš Leonardis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1528-1537

We present a fresh perspective on shot noise corrupted images and noise removal. By viewing image formation as the sequential accumulation of photons on a detector grid, we show that a network trained to predict where the next photon could arrive is in fact solving the minimum mean square error (MMSE) denoising task. This new perspective allows us to make three contributions: (i) We present a new strategy for self-supervised denoising. (ii) We present a new method for sampling from the posterior of possible solutions by iteratively sampling and adding small numbers of photons to the image. (iii) We derive a full generative model by starting this process from an empty canvas. We call this approach generative accumulation of photons (GAP). We evaluate our method quantitatively and qualitatively on 4 new fluorescence microscopy datasets, which will be made available to the community. We find that it outperforms its baselines or performs on-par.

****

Ordinal Classification With Distance Regularization for Robust Brain Age Prediction

Jay Shah, Md Mahfuzur Rahman Siddiquee, Yi Su, Teresa Wu, Baoxin Li; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7882-7891

Age is one of the major known risk factors for Alzheimer's Disease (AD). Detecting AD early is crucial for effective treatment and preventing irreversible brain damage. Brain age, a measure derived from brain imaging reflecting structural changes due to aging, may have the potential to identify AD onset, assess disease risk, and plan targeted interventions. Deep learning-based regression techniques to predict brain age from magnetic resonance imaging (MRI) scans have shown great accuracy recently. However, these methods are subject to an inherent regression to the mean effect, which causes a systematic bias resulting in an overestimation of brain age in young subjects and underestimation in old subjects. This weakens the reliability of predicted brain age as a valid biomarker for downstream clinical applications. Here, we reformulate the brain age prediction task from regression to classification to address the issue of systematic bias. Recognizi ng the importance of preserving ordinal information from ages to understand aging trajectory and monitor aging longitudinally, we propose a novel ORdinal Distance Encoded Regularization (ORDER) loss that incorporates the order of age labels, enhancing the model's ability to capture age-related patterns. Extensive experiments and ablation studies demonstrate that this framework reduces systematic bias, outperforms state-of-art methods by statistically significant margins, and can better capture subtle differences between clinical groups in an independent AD dataset. Our implementation is publicly available at https://github.com/jaygshah/Robust-Brain-Age-Prediction.

****

The Growing Strawberries Dataset: Tracking Multiple Objects With Biological Development Over an Extended Period

Junhan Wen, Camiel R. Verschoor, Chengming Feng, Irina-Mona Epure, Thomas Abeel, Mathijs de Weerdt; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7104-7114

Multiple Object Tracking (MOT) is a rapidly developing research field that targets precise and reliable tracking of objects. Unfortunately, most available MOT datasets typically contain short video clips only, disregarding the indispensable requirement for adequately capturing substantial long-term variations in real-world scenarios. Long-term MOT poses unique challenges due to changes in both the objects and the environment, which remain relatively unexplored. To fill the gap, we propose a time-lapse image dataset inspired by the growth monitoring of strawberries, dubbed "The Growing Strawberries" Dataset (GSD). The data was captured hourly by six cameras, covering a span of 16 months in 2021 and 2022. During this time, it encompassed a total of 24 plants in two separate greenhouses. The changes in appearance, weight, and position during the ripening process, along with variations in the illumination during data collection, distinguish the task from previous MOT research. These practical issues resulted in a drastic performance downgrade in the track identification and association tasks of state-of-the-art MOT algorithms. We believe "The Growing Strawberries" will provide a platform for evaluating such long-term MOT tasks and inspire future research. The dataset is available at https://doi.org/10.4121/e3b31ece-cc88-4638-be10-8ccdd4c5f2f7.v1.

****

Face Presentation Attack Detection by Excavating Causal Clues and Adapting Embedding Statistics

Meiling Fang, Naser Damer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6269-6279

Recent face presentation attack detection (PAD) leverages domain adaptation (DA) and domain generalization (DG) techniques to address performance degradation on unknown domains. However, DA-based PAD methods require access to unlabeled target data, while most DG-based PAD solutions rely on a priori, i.e., known domain labels. Moreover, most DA-/DG-based methods are computationally intensive, demanding complex model architectures and/or multi-stage training processes. This paper proposes to model face PAD as a compound DG task from a causal perspective, linking it to model optimization. We excavate the causal factors hidden in the high-level representation via counterfactual intervention. Moreover, we introduce a class-guided MixStyle to enrich feature-level data distribution within classes instead of focusing on domain information. Both class-guided MixStyle and counterfactual intervention components introduce no extra trainable parameters and negligible computational resources. Extensive cross-dataset and analytic experiments demonstrate the effectiveness and efficiency of our method compared to state-of-the-art PADs. The implementation and the trained weights are publicly available.

****

Glance To Count: Learning To Rank With Anchors for Weakly-Supervised Crowd Counting

Zheng Xiong, Liangyu Chai, Wenxi Liu, Yongtuo Liu, Sucheng Ren, Shengfeng He; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 343-352

Crowd image is arguably one of the most laborious data to annotate. In this paper, we devote to reduce the massive demand of densely labeled crowd data, and propose a novel weakly-supervised setting, in which we leverage the binary ranking of two images with highcontrast crowd counts as training guidance. To enable training under this new setting, we convert the crowd count regression problem to a ranking potential prediction problem. In particular, we tailor a Siamese Ranking Network that predicts the potential scores of two images indicating the ordering of the counts. Hence, the ultimate goal is to assign appropriate potentials for all the crowd images to ensure their orderings obey the ranking labels. On the other hand, potentials reveal the relative crowd sizes but cannot yield an exact crowd count. We resolve this problem by introducing "anchors" during the inference stage. Concretely, anchors are a few images with count labels used for referencing the corresponding counts from potential scores by a simple linear mapping function. We conduct extensive experiments to study various combinations of supervision, and we show that the proposed method outperforms existing weakly-supervised methods without additional labeling effort by a large margin.

****

Gradient-Guided Knowledge Distillation for Object Detectors

Qizhen Lan, Qing Tian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 424-433

Deep learning models have demonstrated remarkable success in object detection, yet their complexity and computational intensity pose a barrier to deploying them in real-world applications (e.g., self-driving perception). Knowledge Distillation (KD) is an effective way to derive efficient models. However, only a small number of KD methods tackle object detection. Also, most of them focus on mimicking the plain features of the teacher model but rarely consider how the features contribute to the final detection. In this paper, we propose a novel approach for knowledge distillation in object detection, named Gradient-guided Knowledge Distillation (GKD). Our GKD uses gradient information to identify and assign more weights to features that significantly impact the detection loss, allowing the student to learn the most relevant features from the teacher. Furthermore, we present bounding-box-aware multi-grained feature imitation (BMFI) to further improve the KD performance. Experiments on the KITTI and COCO Traffic datasets demonstrate our method's efficacy in knowledge distillation for object detection. On one-stage and two-stage detectors, our GKD-BMFI leads to an average of 5.1% and 3.8% mAP improvement, respectively, beating various state-of-the-art KD methods. Our codes are available at: https://github.com/lanqz7766/GKD.

****

SciOL and MuLMS-Img: Introducing a Large-Scale Multimodal Scientific Dataset and Models for Image-Text Tasks in the Scientific Domain

Tim Tarsi, Heike Adel, Jan Hendrik Metzen, Dan Zhang, Matteo Finco, Annemarie Friedrich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4560-4571

In scientific publications, a substantial part of the information is expressed via figures containing images and diagrams. Hence, the retrieval of relevant figures given a natural language query is an important real-world task. However, due to the lack of training and evaluation data, most existing approaches are either limited to one modality or focus on non-scientific domains, making their application to scientific publications challenging. In this paper, we address this gap by introducing two novel datasets: (1) SciOL, the largest openly-licensed pre-training corpus for multimodal models in the scientific domain, covering multiple sciences including materials science, physics, and computer science, and (2) MuLMS-Img, a high-quality dataset in the materials science domain, manually annotated for various image-text tasks. Our experiments show that pre-training large-scale vision-language models on SciOL increases performance considerably across a broad variety of image-text tasks including figure type classification, optical character recognition, captioning, and figure retrieval. Using MuLMS-Img, we show that integrating text-based features extracted via a fine-tuned model for a specific domain can boost cross-modal scientific figure retrieval performance by up to 50%.

****

Few-Shot Event Classification in Images Using Knowledge Graphs for Prompting

Golsa Tahmasebzadeh, Matthias Springstein, Ralph Ewerth, Eric Müller-Budack; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7286-7295

Event classification in images plays a vital role in multimedia analysis especially with the prevalence of fake news on social media and the Web. The majority of approaches for event classification rely on large sets of labeled training data. However, image labels for fine-grained event instances (e.g., 2016 Summer Olympics) can be sparse, incorrect, ambiguous, etc. A few approaches have addressed the lack of labeled data for event classification but cover only few events. Moreover, vision-language models that allow for zero-shot and few-shot classification with prompting have not yet been extensively exploited. In this paper, we propose four different techniques to create hard prompts including knowledge graph information from Wikidata and Wikipedia as well as an ensemble approach for zero-shot event classification. We also integrate prompt learning for state-of-the-art vision-language models to address few-shot event classification. Experimental results on six benchmarks including a new dataset comprising event instances from various domains, such as politics and natural disasters, show that our proposed approaches require much fewer training images than supervised baselines and the state-of-the-art while achieving better results.

****

Simple Token-Level Confidence Improves Caption Correctness

Suzanne Petryk, Spencer Whitehead, Joseph E. Gonzalez, Trevor Darrell, Anna Rohrbach, Marcus Rohrbach; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5742-5752

The ability to judge whether a caption correctly describes an image is a critical part of vision-language understanding. However, state-of-the-art models often misinterpret the correctness of fine-grained details, leading to errors in outputs such as hallucinating objects in generated captions or poor compositional reasoning. In this work, we explore Token-Level Confidence, or TLC, as a simple yet surprisingly effective method to assess caption correctness. Specifically, we fine-tune a vision-language model on image captioning, input an image and proposed caption to the model, and aggregate either algebraic or learned token confidences over words or sequences to estimate image-caption consistency. Compared to sequence-level scores from pretrained models, TLC with algebraic confidence more than doubles image and group scores for compositional reasoning on Winoground. When training data are available, a learned confidence estimator provides further improved performance, reducing object hallucination rates in MS COCO Captions by a relative 30% over the original model and setting a new state-of-the-art.

****

Defending Object Detection Models Against Image Distortions

Mark Ofori-Oduro, Maria Amer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3854-3863

Image distortions pose a significant challenge to object detection. To address this issue, our paper introduces a novel data augmentation method that generates new samples resembling the original training images. The new sample exhibits randomly altered pixels based on a pixel distribution obtained from multiple image distortions using kernel density estimation (KDE). The main steps of our method, GSES, are generating distorted versions of each pixel of an original training image, selecting a set of pixels in each version, and then, for each selected pixel, estimating its distribution using KDE and then sampling one pixel from this distribution. By employing this approach, the new samples possess distorted pixels while maintaining a certain degree of similarity to the original image. This degree of similarity is essential to balance the accuracy of object detection models under distorted and clean images. Our approach improves the accuracy of different object detection models under 15 image distortions, such as motion blur, fog, and noise. For example, the average accuracy of YOLOv4 improves by 9.19% and 9.54% across all 15 distortions added to the COCO and PASCAL datasets, respectively. Our method surpasses other defence methods to combat image distortions. O ur ablation and stability studies show why our method performs well. Moreover, we also show that our method can be well used to improve the accuracy of image classification under 15 distortions and cross-domains. Our code is available at https://github.com/moforio/GSES/.

****

Graph Neural Networks for End-to-End Information Extraction From Handwritten Documents

Yessine Khanfir, Marwa Dhiaf, Emna Ghodhbani, Ahmed Cheikh Rouhou, Yousri Kessentini; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 504-512

Automating Information Extraction (IE) from handwritten documents is a challenging task due to the wide variety of handwriting styles, the presence of noise, and the lack of labeled data. In this work, we propose an end-to-end encoder-decoder model, that incorporates transformers and Graph Convolutional Networks (GCN), to jointly perform Handwritten Text Recognition (HTR) and Named Entity Recognition (NER). The proposed architecture is mainly composed of two parts: a Sparse Graph Transformer Encoder (SGTE), to capture efficient representations of input text images while controlling the propagation of information through the model. The SGTE is followed by a transformer decoder enhanced with a GCN that combines the outputs of the last SGTE layer and the Multi-Head Attention (MHA) block to reinforce the alignment of visual features to characters and Named Entity (NE) tags, resulting in more robust learned representations. The proposed model shows promising results and achieves state-of-the-art performance on the IAM dataset, and in the ICDAR 2017 Information Extraction competition using the Esposalles database.

****

Automated Monitoring of Ear Biting in Pigs by Tracking Individuals and Events

Anicetus Odo, Niall McLaughlin, Ilias Kyriazakis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7095-7103

We propose a system for automated monitoring of ear-biting in pigs. Ear-biting presents a welfare challenge to commercial pig farming, leading to injuries and infections that affect animal welfare. We use a computer vision system to detect and track all pigs and ear-biting events. Our goal is to provide early warning of ear-biting to allow quick intervention to improve the health and welfare of commercial farm animals. We compare several different object detection methods for the detection of individual pigs, including an oriented bounding box detector, which is better suited to the accurate detection of pigs from overhead cameras. We track all pigs and all ear-biting events using a specialised two-stage multi-object tracking system. The tracking system is adapted to match the characteristics of each entity being tracked. The tracking system allows the individual pigs involved in an ear-biting incident to be identified, allowing for targeted welfare interventions. We evaluate our complete system on real farm videos and demonstrate that our complete system improves compared to existing ear-biting detection methods.

****

RSMPNet: Relationship Guided Semantic Map Prediction

Jingwen Sun, Jing Wu, Ze Ji, Yu-Kun Lai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 303-312

In semantic navigation, a top-down map with accurate and complete semantic information is vital to subsequent decision-making. However, due to occlusions and limitations of the robot's field of view (FOV), there are often unobserved areas in the top-down maps. To address this problem, recent works have studied semantic map prediction to complete the top-down maps. In this work, we propose to improve map prediction by integrating relational information. We propose RSMPNet, a relationship-guided semantic map prediction network, which makes use of semantic and spatial relationships to predict unobserved areas from accumulated semantic maps. Specifically, we propose a Relationship Reasoning Layer that includes two modules, namely 1) the Semantic Relationship Graph Reasoning Module (SeGRM) to capture the semantic relationship and 2) the Spatial Relationship Graph Reasoning Module (SpGRM) to utilize the spatial relationship. We also design a semantic r elationship enhanced loss to enhance our model to learn semantic relationship information. Experiments show the effectiveness of our proposed network which achieves state-of-the-art performance in semantic map prediction. Our code and datasets are publicly available at https://github.com/jws39/semantic-mapprediction

****

CARE: Counterfactual-Based Algorithmic Recourse for Explainable Pose Correction

Bhat Dittakavi, Bharathi Callepalli, Aleti Vardhan, Sai Vikas Desai, Vineeth N. Balasubramanian; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4902-4911

With increasing popularity of home-based fitness regimen post-pandemic, there has been a growing interest in fitness monitoring solutions. Owing to this, human pose monitoring has gained significant commercial importance in the field of computer vision. Most efforts in the past focused on the task of human pose classification for various applications. In this work, we instead focus on a critical aspect of human pose monitoring that naturally follows from basic pose classification i.e., pose analysis and correction. Specifically, we study human pose correction through the lens of algorithmic recourse. Algorithmic recourse involves a model providing explanations on a how a model arrived at a decision, along with possible actions to drive the model to output a favorable decision. To this end, we develop CARE (Counterfactuals based Algorithmic Recourse for Explainable pose correction), a novel method that uses counterfactual explanations to provide recourse for incorrect human poses, thereby helping a user correct their pose. Experiments on three diverse datasets, including two fitness datasets and one hand gestures dataset, demonstrate the effectiveness and applicability of CARE.

****

Monocular 3D Object Detection With LiDAR Guided Semi Supervised Active Learning

Aral Hekimoglu, Michael Schmidt, Alvaro Marcos-Ramiro; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2346-2355

We propose a novel semi-supervised active learning framework for monocular 3D object detection with LiDAR guidance (MonoLiG), which leverages all modalities of collected data during model development. We utilize LiDAR to guide the data selection and training of monocular 3D detectors without introducing any overhead in the inference phase. During training, we leverage the LiDAR teacher, monocular student cross-modal framework from semi-supervised learning to distill information from unlabeled data as pseudo-labels. To handle the differences in sensor characteristics, we propose a data noise-based weighting mechanism to reduce the effect of propagating noise from LiDAR modality to monocular. For selecting which samples to label to improve the model performance, we propose a sensor consistency-based selection score that is also coherent with the training objective. Extensive experimental results on KITTI and Waymo datasets verify the effectiveness of our proposed framework. In particular, our selection strategy consistently outperforms state-of-the-art active learning baselines, yielding up to 17% better saving rate in labeling costs. Our training strategy attains the top place in KITTI 3D and bird's-eye-view (BEV) monocular object detection official benchmarks by improving the BEV Average Precision (AP) by 2.02. Code is shared at https://github.com/aralhekimoglu/monolig.

****

S3AD: Semi-Supervised Small Apple Detection in Orchard Environments

Robert Johanson, Christian Wilms, Ole Johannsen, Simone Frintrop; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7076-7085

Crop detection is integral for precision agriculture applications such as automated yield estimation or fruit picking. However, crop detection, e.g., apple detection in orchard environments remains challenging due to a lack of large-scale datasets and the small relative size of the crops in the image. In this work, we address these challenges by reformulating the apple detection task in a semi-supervised manner. To this end, we provide the large, high-resolution dataset MAD comprising 105 labeled images with 14,667 annotated apple instances and 4,440 unlabeled images. Utilizing this dataset, we also propose a novel Semi-Supervised S mall Apple Detection system S3AD based on contextual attention and selective tiling to improve the challenging detection of small apples, while limiting the computational overhead. We conduct an extensive evaluation on MAD and the MSU dataset, showing that S3AD substantially outperforms strong fully-supervised baselines, including several small object detection systems, by up to 14.9%. Additionally, we exploit the detailed annotations of our dataset w.r.t. apple properties to analyze the influence of relative size or level of occlusion on the results of various systems, quantifying current challenges.

****

Task-Oriented Human-Object Interactions Generation With Implicit Neural Representations

Quanzhou Li, Jingbo Wang, Chen Change Loy, Bo Dai; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3035-3044

Digital human motion synthesis is a vibrant research field with applications in movies, AR/VR, and video games. Whereas methods were proposed to generate natural and realistic human motions, most only focus on modeling humans and largely ignore object movements. Generating task-oriented human-object interaction motions in simulation is challenging. For different intents of using the objects, humans conduct various motions, which requires the human first to approach the objects and then make them move consistently with the human instead of staying still. Also, to deploy in downstream applications, the synthesized motions are desired to be flexible in length, providing options to personalize the predicted motions for various purposes. To this end, we propose TOHO: Task-Oriented Human-Object Interactions Generation with Implicit Neural Representations, which generates full human-object interaction motions to conduct specific tasks, given only the task type, the object, and a starting human status. TOHO generates human-object motions in four steps: 1) it first estimates the object's final position given the task intent; 2) it then generates keyframe poses grasping the objects; 3) after that, it infills the keyframes and generates continuous motions; 4) finally, it applies a compact closed-form object motion estimation to generate the object motion. Our method generates continuous motions that are parameterized only by the temporal coordinate, which allows for upsampling of the sequence to arbitrary frames and adjusting the motion speeds by designing the temporal coordinate vector. This work takes a step further toward general human-scene interaction simulation.

****

Convolutional Masked Image Modeling for Dense Prediction Tasks on Pathology Images

Yan Yang, Liyuan Pan, Liu Liu, Eric A. Stone; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7798-7808

This paper studies a convolutional masked image modeling approach for boosting downstream dense prediction tasks on pathology images. Our method is self-supervised, and entails two strategies in sequence. Considering features contained in the pathology images usually have a large spatial span, e.g., glands, we insert [MASK] tokens to the masked regions after the stem layer of the convolutional network for encoding unmasked pixels, which facilitates information propagation through masked regions for reconstructing unmasked pixels. Furthermore, the pathology images contain features that are represented in diverse affine shapes and color spaces. We, therefore, enforce the network to learn the affine and color invariant embedding by imposing transformation constraints between the unmasked image-encoded embedding and reconstruction targets. Our approach is simple but effective. With extensive experiments on standard benchmark datasets, we demonstrate superior transfer learning performance on downstream tasks over past state-of-the-art approaches.

****

Controlling Virtual Try-On Pipeline Through Rendering Policies

Kedan Li, Jeffrey Zhang, Shao-Yu Chang, David Forsyth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5866-5875

This paper shows how to impose rendering policies on a virtual try-on (VTON) pip eline. Our rendering policies are lightweight procedural descriptions of how the pipeline should render outfits or render particular types of garments. Our policies are procedural expressions describing offsets to the control points for each set of garment types. The policies are easily authored and are generalizable to any outfit composed of garments of similar types. We describe a VTON pipeline that accepts our policies to modify garment drapes and produce high-quality try-on images with garment attributes preserved. Layered outfits are a particular challenge to VTON systems because learning to coordinate warps between multiple garments so that nothing sticks out is difficult. Our rendering policies offer a lightweight and effective procedure to achieve this coordination, while also allowing precise manipulation of drape. Drape describes the way in which a garment is worn (for example, a shirt could be tucked or untucked). Quantitative and qualitative evaluations demonstrate that our method allows effective manipulation of drape and produces significant measurable improvements in rendering quality for complicated layering interactions.

****

Interpretable Object Recognition by Semantic Prototype Analysis

Qiyang Wan, Ruiping Wang, Xilin Chen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 800-809

People can usually give reasons for recognizing a particular object as a specific category, using various means such as body language (by pointing out) and natural language (by telling). This inspires us to develop a recognition model with such principles to explain the recognition process to enhance human trust. We propose Semantic Prototype Analysis Network (SPANet), an interpretable object recognition approach that enables models to explicate the decision process more lucidly and comprehensibly to humans by "pointing out where to focus" and "telling about why it is" simultaneously. With the proposed method, some part prototypes with semantic concepts will be provided to elaborate on the classification together with a group of visualized samples to achieve both part-wise and semantic interpretability. The results of extensive experiments demonstrate that SPANet is able to recognize objects almost as well as the non-interpretable models, at the same time generating intelligible explanations for its decision process.

****

Assist Is Just As Important as the Goal: Image Resurfacing To Aid Model's Robust Prediction

Abhijith Sharma, Phil Munz, Apurva Narayan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3833-3842

Adversarial patches threaten visual AI models in the real world. The number of patches in a patch attack is variable and determines the attack's potency in a specific environment. Most existing defenses assume a single patch in the scene, and the multiple patch scenario are shown to overcome them. This paper presents a model-agnostic defense against patch attacks based on total variation for image resurfacing (TVR). The TVR is an image-cleansing method that processes images to remove probable adversarial regions. TVR can be utilized solely or augmented with a defended model, providing multi-level security for robust prediction. TVR nullifies the influence of patches in a single image scan with no prior assumption on the number of patches in the scene. We validate TVR on the ImageNet-Patch benchmark dataset and with real-world physical objects, demonstrating its ability to mitigate patch attack.

****

Prompting Classes: Exploring the Power of Prompt Class Learning in Weakly Supervised Semantic Segmentation

Balamurali Murugesan, Rukhshanda Hussain, Rajarshi Bhattacharya, Ismail Ben Ayed, Jose Dolz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 291-302

Recently, CLIP-based approaches have exhibited remarkable performance on generalization and few-shot learning tasks, fueled by the power of contrastive language-vision pre-training. In particular, prompt tuning has emerged as an effective strategy to adapt the pre-trained language-vision models to downstream tasks by employing task-related textual tokens. Motivated by this progress, in this work w e question whether other fundamental problems, such as weakly supervised semantic segmentation (WSSS), can benefit from prompt tuning. Our findings reveal two interesting observations that shed light on the impact of prompt tuning on WSSS. First, modifying only the class token of the text prompt results in a greater impact on the Class Activation Map (CAM), compared to arguably more complex strategies that optimize the context. And second, the class token associated with the image ground truth does not necessarily correspond to the category that yields the best CAM. Motivated by these observations, we introduce a novel approach based on a PrOmpt cLass lEarning (POLE) strategy. Through extensive experiments we demonstrate that our simple, yet efficient approach achieves SOTA performance in a well-known WSSS benchmark. These results highlight not only the benefits of language-vision models in WSSS but also the potential of prompt learning for this problem. The code is available at https://anonymous.4open.science/r/WSS_POLE-DB45/README.md

****

From Denoising Training To Test-Time Adaptation: Enhancing Domain Generalization for Medical Image Segmentation

Ruxue Wen, Hangjie Yuan, Dong Ni, Wenbo Xiao, Yaoyao Wu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 464-474

In medical image segmentation, domain generalization poses a significant challenge due to domain shifts caused by variations in data acquisition devices and other factors. These shifts are particularly pronounced in the most common scenario, which involves only single-source domain data due to privacy concerns. To address this, we draw inspiration from the self-supervised learning paradigm that effectively discourages overfitting to the source domain. We propose the Denoising Y-Net (DeY-Net), a novel approach incorporating an auxiliary denoising decoder into the basic U-Net architecture. The auxiliary decoder aims to perform denoising training, augmenting the domain-invariant representation that facilitates domain generalization. Furthermore, this paradigm provides the potential to utilize unlabeled data. Building upon denoising training, we propose Denoising Test Time Adaptation (DeTTA) that further: (i) adapts the model to the target domain in a sample-wise manner, and (ii) adapts to the noise-corrupted input. Extensive experiments conducted on widely-adopted liver segmentation benchmarks demonstrate significant domain generalization improvements over our baseline and state-of-the-art results compared to other methods. Code is available at https://github.com/WenRuxue/DeTTA.

****

Implicit Neural Image Stitching With Enhanced and Blended Feature Reconstruction

Minsu Kim, Jaewon Lee, Byeonghun Lee, Sunghoon Im, Kyong Hwan Jin; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4087-4096

Existing frameworks for image stitching often provide visually reasonable stitchings. However, they suffer from blurry artifacts and disparities in illumination, depth level, etc. Although the recent learning-based stitchings relax such disparities, the required methods impose sacrifice of image qualities failing to capture high-frequency details for stitched images. To address the problem, we propose a novel approach, implicit Neural Image Stitching (NIS) that extends arbitrary-scale super-resolution. Our method estimates Fourier coefficients of images for quality-enhancing warps. Then, the suggested model blends color mismatches and misalignment in the latent space and decodes the features into RGB values of stitched images. Our experiments show that our approach achieves improvement in resolving the low-definition imaging of the previous deep image stitching with favorable accelerated image-enhancing methods. Our source code is available at https://github.com/minshu-kim/NIS.

****

360BEV: Panoramic Semantic Mapping for Indoor Bird's-Eye View

Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, Rainer Stiefelhagen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 373-382

Seeing only a tiny part of the whole is not knowing the full circumstance. Bird's-eye-view (BEV) perception, a process of obtaining allocentric maps from egocentric views, is restricted when using a narrow Field of View (FoV) alone. In this work, mapping from 360deg panoramas to BEV semantics, the 360BEV task, is established for the first time to achieve holistic representations of indoor scenes in a top-down view. Instead of relying on narrow-FoV image sequences, a panoramic image with depth information is sufficient to generate a holistic BEV semantic map. To benchmark 360BEV, we present two indoor datasets, 360BEV-Matterport and 360BEV-Stanford, both of which include egocentric panoramic images and semantic segmentation labels, as well as allocentric semantic maps. Besides delving deep into different mapping paradigms, we propose a dedicated solution for panoramic semantic mapping, namely 360Mapper. Through extensive experiments, our methods achieve 44.32% and 45.78% mIoU on both datasets respectively, surpassing previous counterparts with gains of +7.60% and +9.70% in mIoU.

****

Semi-Supervised Semantic Depth Estimation Using Symbiotic Transformer and NearFarMix Augmentation

Md Awsafur Rahman, Shaikh Anowarul Fattah; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 250-259

In computer vision, depth estimation is crucial for domains like robotics, autonomous vehicles, augmented reality, and virtual reality. Integrating semantics with depth enhances scene understanding through reciprocal information sharing. However, the scarcity of semantic information in datasets poses challenges. Existing convolutional approaches with limited local receptive fields hinder the full utilization of the symbiotic potential between depth and semantics. This paper introduces a dataset-invariant semi-supervised strategy to address the scarcity of semantic information. It proposes the Depth Semantics Symbiosis module, leveraging the Symbiotic Transformer for achieving comprehensive mutual awareness by information exchange within both local and global contexts. Additionally, a novel augmentation, NearFarMix is introduced to combat overfitting and compensate both depth-semantic tasks by strategically merging regions from two images, generating diverse and structurally consistent samples with enhanced control. Extensive experiments on NYU-Depth-V2 and KITTI datasets demonstrate the superiority of our proposed techniques in indoor and outdoor environments.

****

Query-Guided Attention in Vision Transformers for Localizing Objects Using a Single Sketch

Aditay Tripathi, Anand Mishra, Anirban Chakraborty; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1083-1092

In this study, we explore sketch-based object localization on natural images. Given a crude hand-drawn object sketch, the task is to locate all instances of that object in the target image. This problem proves difficult due to the abstract nature of hand-drawn sketches, variations in the style and quality of sketches, and the large domain gap between the sketches and the natural images. Existing solutions address this using attention-based frameworks to merge query information into image features. Yet, these methods often integrate query features after independently learning image features, causing inadequate alignment and as a result incorrect localization. In contrast, we propose a novel sketch-guided vision transformer encoder that uses cross-attention after each block of the transformer-based image encoder to learn query-conditioned image features, leading to stronger alignment with the query sketch. Further, at the decoder's output, object and sketch features are refined better to align the representation of objects with the sketch query, thereby improving localization. The proposed model also generalizes to the object categories not seen during training, as the target image features learned by the proposed model are query-aware. Our framework can utilize multiple sketch queries via a trainable novel sketch fusion strategy. The model is evaluated on the images from the public benchmark, MS-COCO, using the sketch queries from QuickDraw! and Sketchy datasets. Compared with existing localization methods, the proposed approach gives a 6.6% and 8.0% improvement in mAP for seen objects using sketch queries from QuickDraw! and Sketchy datasets, respectiv ely, and a 12.2% improvement in AP@50 for large objects that are 'unseen' during training.

****

I-AI: A Controllable & Interpretable AI System for Decoding Radiologists' Intense Focus for Accurate CXR Diagnoses

Trong Thang Pham, Jacob Brecheisen, Anh Nguyen, Hien Nguyen, Ngan Le; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7850-7859

In the field of chest X-ray (CXR) diagnosis, existing works often focus solely on determining where a radiologist looks, typically through tasks such as detection, segmentation, or classification. However, these approaches are often designed as black-box models, lacking interpretability. In this paper, we introduce Interpretable Artificial Intelligence (I-AI) a novel and unified controllable interpretable pipeline for decoding the intense focus of radiologists in CXR diagnosis. Our I-AI addresses three key questions: where a radiologist looks, how long they focus on specific areas, and what findings they diagnose. By capturing the intensity of the radiologist's gaze, we provide a unified solution that offers insights into the cognitive process underlying radiological interpretation. Unlike current methods that rely on black-box machine learning models, which can be prone to extracting erroneous information from the entire input image during the diagnosis process, we tackle this issue by effectively masking out irrelevant information. Our proposed I-AI leverages a vision-language model, allowing for precise control over the interpretation process while ensuring the exclusion of irrelevant features. To train our I-AI model, we utilize an eye gaze dataset to extract anatomical gaze information and generate ground truth heatmaps. Through extensive experimentation, we demonstrate the efficacy of our method. We showcase that the attention heatmaps, designed to mimic radiologists' focus, encode sufficient and relevant information, enabling accurate classification tasks using only a portion of CXR. The code, checkpoints, and data are at https://github.com/UARK-AICV/IAI.

****

Diffuse and Restore: A Region-Adaptive Diffusion Model for Identity-Preserving Blind Face Restoration

Maitreya Suin, Nithin Gopalakrishnan Nair, Chun Pong Lau, Vishal M. Patel, Rama Chellappa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6343-6352

Blind face restoration (BFR) from severely degraded face images in the wild is a highly ill-posed problem. Due to the complex unknown degradation, existing generative works typically struggle to restore realistic details when the input is of poor quality. Recently, diffusion-based approaches were successfully used for high-quality image synthesis. But, for BFR, maintaining a balance between the fidelity of the restored image and the reconstructed identity information is important. Minor changes in certain facial regions may alter the identity or degrade the perceptual quality. With this observation, we present a conditional diffusion-based framework for BFR. We alleviate the drawbacks of existing diffusion-based approaches and design an region-adaptive strategy. Specifically, we use a identity preserving conditioner network to recover the identity information from the input image as much as possible and use that to guide the reverse diffusion process, specifically for important facial locations that contribute the most to the identity. This leads to a significant improvement in perceptual quality as well as face-recognition scores over existing GAN and diffusion-based restoration models. Our approach achieves superior results to prior art on a range of real and synthetic datasets, particularly for severely degraded face images.

****

Interaction Region Visual Transformer for Egocentric Action Anticipation

Debaditya Roy, Ramanathan Rajendiran, Basura Fernando; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6740-6750

Human-object interaction (HOI) and temporal dynamics along the motion paths are the most important visual cues for egocentric action anticipation. Especially, i nteraction regions covering objects and the human hand reveal significant visual cues to predict future human actions. However, how to incorporate and capture these important visual cues in modern video Transformer architecture remains a challenge, especially because integrating inductive biases into Transformers is hard. We leverage the effective MotionFormer that models motion dynamics to incorporate interaction regions using spatial cross-attention and further infuse contextual information using trajectory cross-attention to obtain an interaction-centric video representation for action anticipation. We term our model InAViT which achieves state-of-the-art action anticipation performance on large-scale egocentric datasets EPICKTICHENS100 (EK100) and EGTEA Gaze+. On the EK100 evaluation server, InAViT is on top of the public leader board (at the time of submission) where it outperforms the second-best model by 3.3% on mean-top5 recall. We will release the code.

****

Preserving Image Properties Through Initializations in Diffusion Models

Jeffrey Zhang, Shao-Yu Chang, Kedan Li, David Forsyth; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5242-5250

Retail photography imposes specific requirements on images. For instance, images may need uniform background colors, consistent model poses, centered products, and consistent lighting. Minor deviations from these standards impact a site's aesthetic appeal, making the images unsuitable for use. We show that Stable Diffusion methods, as currently applied, do not respect these requirements. The usual practice of training the denoiser with a very noisy image and starting inference with a sample of pure noise leads to inconsistent generated images during inference. This inconsistency occurs because it is easy to tell the difference between samples of the training and inference distributions. As a result, a network trained with centered retail product images with uniform backgrounds generates images with erratic backgrounds. The problem is easily fixed by initializing inference with samples from an approximation of noisy images. However, in using such an approximation, the joint distribution of text and noisy image at inference time still slightly differs from that at training time. This discrepancy is corrected by training the network with samples from the approximate noisy image distribution. Extensive experiments on real application data show significant qualitative and quantitative improvements in performance from adopting these procedures. Finally, our procedure can interact well with other control-based methods to further enhance the controllability of diffusion-based methods.

****

Limited Data, Unlimited Potential: A Study on ViTs Augmented by Masked Autoencoders

Srijan Das, Tanmay Jain, Dominick Reilly, Pranav Balaji, Soumyajit Karmakar, Shyam Marjit, Xiang Li, Abhijit Das, Michael S. Ryoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6878-6888

Vision Transformers (ViTs) have become ubiquitous in computer vision. Despite their success, ViTs lack inductive biases, which can make it difficult to train them with limited data. To address this challenge, prior studies suggest training ViTs with self-supervised learning (SSL) and fine-tuning sequentially. However, we observe that jointly optimizing ViTs for the primary task and a Self-Supervised Auxiliary Task (SSAT) is surprisingly beneficial when the amount of training data is limited. We explore the appropriate SSL tasks that can be optimized alongside the primary task, the training schemes for these tasks, and the data scale at which they can be most effective. Our findings reveal that SSAT is a powerful technique that enables ViTs to leverage the unique characteristics of both the self-supervised and primary tasks, achieving better performance than typical ViT pre-training with SSL and fine-tuning sequentially. Our experiments, conducted on 10 datasets, demonstrate that SSAT significantly improves ViT performance while reducing carbon footprint. We also confirm the effectiveness of SSAT in the video domain for deepfake detection, showcasing its generalizability.

****

Unsupervised Co-Generation of Foreground-Background Segmentation From Text-to-Im age Synthesis

Yeruru Asrar Ahmed, Anurag Mittal; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5058-5069

Text-to-Image (T2I) synthesis is a challenging task requiring modelling both textual and image domains and their relationship. The substantial improvement in image quality achieved by recent works has paved the way for numerous applications such as language-aided image editing, computer-aided design, text-based image retrieval, and training data augmentation. In this work, we ask a simple question: Along with realistic images, can we obtain any useful by-product (e.g., foreground / background or multi-class segmentation masks, detection labels) in an unsupervised way that will also benefit other computer vision tasks and applications?. In an attempt to answer this question, we explore generating realistic images and their corresponding foreground / background segmentation masks from the given text. To achieve this, we experiment the concept of co-segmentation along with GAN. Specifically, a novel GAN architecture called Co-Segmentation Inspired GAN (COS-GAN) is proposed that generates two or more images simultaneously from different noise vectors and utilises a spatial co-attention mechanism between the image features to produce realistic segmentation masks for each of the generated images. The advantages of such an architecture are two-fold: 1) The generated segmentation masks can be used to focus on foreground and background exclusively to improve the quality of generated images, and 2) the segmentation masks can be used as a training target for other tasks, such as object localisation and segmentation. Extensive experiments conducted on CUB, Oxford-102, and COCO datasets show that COS-GAN is able to improve visual quality and generate reliable foreground / background masks for the generated images.

****

High-Fidelity Zero-Shot Texture Anomaly Localization Using Feature Correspondence Analysis

Andrei-Timotei Ardelean, Tim Weyrich; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1134-1144

We propose a novel method for Zero-Shot Anomaly Localization on textures. The task refers to identifying abnormal regions in an otherwise homogeneous image. To obtain a high-fidelity localization, we leverage a bijective mapping derived from the 1-dimensional Wasserstein Distance. As opposed to using holistic distances between distributions, the proposed approach allows pinpointing the non-conformity of a pixel in a local context with increased precision. By aggregating the contribution of the pixel to the errors of all nearby patches we obtain a reliable anomaly score estimate. We validate our solution on several datasets and obtain more than a 40% reduction in error over the previous state of the art on the MVTec AD dataset in a zero-shot setting. Also see https://reality.tf.fau.de/pub/ardelean2024highfidelity.html.

****

Hierarchical Text Spotter for Joint Text Spotting and Layout Analysis

Shangbang Long, Siyang Qin, Yasuhisa Fujii, Alessandro Bissacco, Michalis Raptis; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 903-913

We propose Hierarchical Text Spotter (HTS), a novel method for the joint task of word-level text spotting and geometric layout analysis. HTS can recognize text in an image and identify its 4-level hierarchical structure: characters, words, lines, and paragraphs. The proposed HTS is characterized by two novel components: (1) a Unified-Detector-Polygon (UDP) that produces Bezier Curve polygons of text lines and an affinity matrix for paragraph grouping between detected lines; (2) a Line-to-Character-to-Word (L2C2W) recognizer that splits lines into characters and further merges them back into words. HTS achieves state-of-the-art results on multiple word-level text spotting benchmark datasets as well as geometric layout analysis tasks.

****

Label-Free Synthetic Pretraining of Object Detectors

Hei Law, Jia Deng; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 946-956

We propose a new approach, Synthetic Optimized Layout with Instance Detection (SOLID), to pretrain object detectors with synthetic images. Our "SOLID" approach consists of two main components: (1) generating synthetic images using a collection of unlabelled 3D models with optimized scene arrangement; (2) pretraining an object detector on "instance detection" task - given a query image depicting an object, detecting all instances of the exact same object in a target image. Our approach does not need any semantic labels for pretraining and allows the use of arbitrary, diverse 3D models. Experiments on COCO show that with optimized data generation and a proper pretraining task, synthetic data can be highly effective data for pretraining object detectors. In particular, pretraining on rendered images achieves performance competitive with pretraining on real images while using significantly less computing resources.

****

Tracking Tiny Insects in Cluttered Natural Environments Using Refinable Recurrent Neural Networks

Lars Haalck, Sebastian Thiele, Benjamin Risse; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7126-7135

Visual tracking of tiny and low-contrast objects such as insects in cluttered natural environments is a very challenging computer vision task. This is particularly true for machine learning algorithms, which usually require distinct visual foreground features to reliably identify the object of interest. Here, we propose a novel deep learning-based tracking framework capable of detecting tiny and visually camouflaged ants (covering only a few pixels) in complex and dynamic high-resolution videos. In particular, we introduce refinable recurrent Hourglass Networks, which combine colour and temporal information to continuously detect insects recorded using a freely moving camera. Moreover, this architecture provides comprehensible heatmaps of positional estimations and a seamless integration of optional user-input to further refine the tracking results if necessary. We evaluated our algorithm on an extremely challenging wildlife ant dataset with a resolution of 1024x1024 and report a mean deviation of 19 pixels from the ground truth (object 30 px) without any user input. By providing only 0.6% manual locations this accuracy can be improved to a mean deviation of 9 pixels. A comparison to a well known deep learning-based single frame detection algorithm (YOLOv7), two state-of-the-art tracking methods (ToMP and KeepTrack), a probabilistic tracking framework and a comprehensive ablation study reveal superior performances in all our experiments. Our tracking framework therefore provides a foundation for challenging tiny single-object tracking scenarios and a practical and interactive solution for biologists and ecologists.

****

RGB-X Object Detection via Scene-Specific Fusion Modules

Sri Aditya Deevi, Connor Lee, Lu Gan, Sushruth Nagesh, Gaurav Pandey, Soon-Jo Chung; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7366-7375

Multimodal deep sensor fusion has the potential to enable autonomous vehicles to visually understand their surrounding environments in all weather conditions. However, existing deep sensor fusion methods usually employ convoluted architectures with intermingled multimodal features, requiring large coregistered multimodal datasets for training. In this work, we present an efficient and modular RGB-X fusion network that can leverage and fuse pretrained single-modal models via scene-specific fusion modules, thereby enabling joint input-adaptive network architectures to be created using small, coregistered multimodal datasets. Our experiments demonstrate the superiority of our method compared to existing works on RGB-thermal and RGB-gated datasets, performing fusion using only a small amount of additional parameters. Our code is available at https://github.com/dsriaditya999/RGBXFusion.

****

3D-Aware Talking-Head Video Motion Transfer

Haomiao Ni, Jiachen Liu, Yuan Xue, Sharon X. Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4954-4964

Motion transfer of talking-head videos involves generating a new video with the appearance of a subject video and the motion pattern of a driving video. Current methodologies primarily depend on a limited number of subject images and 2D representations, thereby neglecting to fully utilize the multi-view appearance features inherent in the subject video. In this paper, we propose a novel 3D-aware talking-head video motion transfer network, Head3D, which fully exploits the subject appearance information by generating a visually-interpretable 3D canonical head from the 2D subject frames with a recurrent network. A key component of our approach is a self-supervised 3D head geometry learning module, designed to predict head poses and depth maps from 2D subject video frames. This module facilitates the estimation of a 3D head in canonical space, which can then be transformed to align with driving video frames. Additionally, we propose an attention-based fusion network to combine the background and other details from subject frames with the 3D subject head to produce the synthetic target video. Our extensive experiments on two public talking-head video datasets demonstrate that Head3D outperforms both 2D and 3D prior arts in the practical cross-identity setting, with evidence showing it can be readily adapted to the pose-controllable novel view synthesis task.

****

Lightweight Thermal Super-Resolution and Object Detection for Robust Perception in Adverse Weather Conditions

Pranjay Shyam, HyunJin Yoo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7471-7482

In this work, we examine the potential application of thermal cameras in improving perception capabilities in adverse weather conditions like snow, night-time driving, and haze, focusing on retaining the performance of Advanced Driver Assistance Systems (ADAS), thus enhancing its functionality and safety characteristics. While thermal sensors offer the advantage of robust information capture in adverse weather conditions, their integration is plagued with issues surrounding poor feature capture in normal conditions, low imaging resolution, and high sensor costs. We address the former by formulating the problem definition as information switching wherein thermal images are selected when visible images are degraded. Furthermore, we consider a single object detector for RGB and thermal images to ensure low latency. We propose utilizing a learnable projection function that translates the thermal image into RGB color space, thus providing minimal modifications to the underlying object detector. We address the issues of low imaging resolution and cost by proposing a novel procedure that combines super-resolution and object detection, enabling the utilization of low-resolution and low-cost uncooled thermal imaging sensors. To ensure the complete pipeline meets the actual deployment requirements of real-time inference on resource-constrained devices, we introduce a lightweight super-resolution algorithm, implementing optimizations within the network structure followed by global pruning. In addition, to improve the feature representations extracted by lightweight encoders, we propose a bidirectional feature pyramid network to enhance the feature representation. We demonstrate the efficacy of the proposed mechanism through extensive simulated evaluations on automotive datasets such as FLIR, KAIST, DENSE, and Freiburg Thermal.

****

Revolutionize the Oceanic Drone RGB Imagery With Pioneering Sun Glint Detection and Removal Techniques

Jiangying Qin, Ming Li, Jie Zhao, Jiageng Zhong, Hanqi Zhang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8326-8335

The issue of sun glint poses a significant challenge for ocean remote sensing with high-resolution ocean drone imagery, as it contaminates images and obstructs crucial features in shallow-waters, leading to inaccurate benthic substrates identification. While various physics-based statistical solutions have been proposed to address this optical issue in remote sensing, there is a lack of sun glint detection and removal methods specifically designed for high-resolution consumer-grade drone RGB imagery. In this paper, we present a pioneering pipeline for sun glint detection and removal in high-resolution drone RGB images, aiming to res tore the real features that are hindered by sun glint. Our approach involves the development of a Foreground Attention-based Semantic Segmentation Network (FANet) for accurate and precise sun glint detection, while effective sun glint removal is achieved through pixel propagation using an optical flow field. Experimental results demonstrate the effectiveness of our FANet in identifying sun glint, achieving IoU accuracy of 81.34% for sun glint pixels and 99.52% for non-sun glint background pixels. Furthermore, the quantitative evaluation of sun glint removal using two well-known metrics show that our method outperforms the GAN-based image restoration method (DeepFillv2) and the conventional image interpolation method (Fast Marching Method, hereafter referred to as FMM). Thus, our pipeline lays the foundation for accurate and precise marine costal ecological monitoring and seafloor topographic mapping using consumer-grade drone at a low cost.

****

Revisiting Pixel-Level Contrastive Pre-Training on Scene Images

Zongshang Pang, Yuta Nakashima, Mayu Otani, Hajime Nagahara; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1784-1793

Contrastive image representation learning through instance discrimination has shown impressive transfer performance. Recent strategies have focused on pushing the limit of their transfer performance for dense prediction tasks, particularly when conducting pre-training on scene images with complex structures. Initial approaches employ pixel-level contrastive pre-training to optimize dense spatial features, while subsequent methods utilize region-mining algorithms to capture holistic regional semantics and address the issue of semantically inconsistent scene image crops. In this paper, we revisit pixel-level contrastive pre-training on scene images. Contrary to the assumption that pixel-level learning falls short in achieving these objectives, we demonstrate its under-explored potentials: (1) it can effectively learn holistic regional semantics more simply compared to region-level methods, and (2) it intrinsically provides tools to mitigate the impact of semantically inconsistent views involved with scene-level training images. We propose PixCon, a pixel-level contrastive learning framework, and explore two variants with different positive matching strategies to investigate the potential of pixel-level learning. Additionally, when PixCon incorporates a novel semantic reweighting approach tailored for scene image pre-training, it outperforms or matches the performance of previous region-level methods in object detection and semantic segmentation tasks across multiple benchmarks.

****

Neural Textured Deformable Meshes for Robust Analysis-by-Synthesis

Angtian Wang, Wufei Ma, Alan Yuille, Adam Kortylewski; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3108-3117

Human vision demonstrates higher robustness than current AI algorithms under out-of-distribution scenarios. It has been conjectured such robustness benefits from performing analysis-by-synthesis. Our paper formulates triple vision tasks in a consistent manner using approximate analysis-by-synthesis by render-and-compare algorithms on neural features. In this work, we introduce Neural Textured Deformable Meshes, which involve the object model with deformable geometry that allows optimization on both camera parameters and object geometries. The deformable mesh is parameterized as a neural field, and covered by whole-surface neural texture maps, which are trained to have spatial discriminability. During inference, we extract the feature map of the test image and subsequently optimize the 3D pose and shape parameters of our model using differentiable rendering to best reconstruct the target feature map. We show that our analysis-by-synthesis is much more robust than conventional neural networks when evaluated on real-world images and even in challenging out-of-distribution scenarios, such as occlusion and domain shift. Our algorithms are competitive with standard algorithms when tested on conventional performance measures.

****

PsyMo: A Dataset for Estimating Self-Reported Psychological Traits From Gait

Adrian Cosma, Emilian Radoi; Proceedings of the IEEE/CVF Winter Conference on Ap plications of Computer Vision (WACV), 2024, pp. 4603-4613

Psychological trait estimation from external factors such as movement and appearance is a challenging and long-standing problem in psychology, and is principally based on the psychological theory of embodiment. To date, attempts to tackle this problem have utilized private small-scale datasets with intrusive body-attached sensors. Potential applications of an automated system for psychological trait estimation include estimation of occupational fatigue and psychology, and marketing and advertisement. In this work, we propose PsyMo (Psychological traits from Motion), a novel, multi-purpose and multi-modal dataset for exploring psychological cues manifested in walking patterns. We gathered walking sequences from 312 subjects in 7 different walking variations and 6 camera angles. In conjunction with walking sequences, participants filled in 6 psychological questionnaires, totaling 17 psychometric attributes related to personality, self-esteem, fatigue, aggressiveness and mental health. We propose two evaluation protocols for psychological trait estimation. Alongside the estimation of self-reported psychological traits from gait, the dataset can be used as a drop-in replacement to benchmark methods for gait recognition. We anonymize all cues related to the identity of the subjects and publicly release only silhouettes, 2D / 3D human skeletons and 3D SMPL human meshes.

****

Back to Optimization: Diffusion-Based Zero-Shot 3D Human Pose Estimation

Zhongyu Jiang, Zhuoran Zhou, Lei Li, Wenhao Chai, Cheng-Yen Yang, Jenq-Neng Hwang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6142-6152

Learning-based methods have dominated the 3D human pose estimation (HPE) tasks with significantly better performance in most benchmarks than traditional optimization-based methods. Nonetheless, 3D HPE in the wild is still the biggest challenge of learning-based models, whether with 2D-3D lifting, image-to-3D, or diffusion-based methods, since the trained networks implicitly learn camera intrinsic parameters and domain-based 3D human pose distributions and estimate poses by statistical average. On the other hand, the optimization-based methods estimate results case-by-case, which can predict more diverse and sophisticated human poses in the wild. By combining the advantages of optimization-based and learning-based methods, we propose the Zero-shot Diffusion-based Optimization (ZeDO) pipeline for 3D HPE to solve the problem of cross-domain and in-the-wild 3D HPE. Our multi-hypothesis ZeDO achieves state-of-the-art (SOTA) performance on Human3.6M as minMPJPE 51.4mm without training with any 2D-3D or image-3D pairs. Moreover, our single-hypothesis ZeDO achieves SOTA performance on 3DPW dataset with PA-MPJPE 42.6mm on cross-dataset evaluation, which even outperforms learning-based methods trained on 3DPW.

****

IKEA Ego 3D Dataset: Understanding Furniture Assembly Actions From Ego-View 3D Point Clouds

Yizhak Ben-Shabat, Jonathan Paul, Eviatar Segev, Oren Shrout, Stephen Gould; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4355-4364

We propose a novel dataset for ego-view 3D point cloud action recognition. While there has been extensive research on understanding human actions in RGB videos in recent years, the exploration of its 3D point cloud counterpart has been relatively limited. Furthermore, RGB ego-view datasets are rapidly growing, however, 3D point cloud ego-view datasets are scarce at best. Existing 3D datasets are limited in several ways, some include actions that are distinguishable by full-body motion while others use a distant static sensor that hinders the recognition of small objects. We introduce a new point cloud action recognition dataset---the IKEA Ego 3D dataset. It includes sequences of point clouds captured from an ego-view using a HoloLens 2 device. The dataset consists of approximately 493k frames and 56 classes of intricate furniture assembly actions of four different furniture types. We evaluate the performance of various state-of-the-art 3D action recognition methods on the proposed dataset and show that it is very challenging.

****

Concept-Centric Transformers: Enhancing Model Interpretability Through Object-Centric Concept Learning Within a Shared Global Workspace

Jinyung Hong, Keun Hee Park, Theodore P. Pavlic; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4880-4891

Many interpretable AI approaches have been proposed to provide plausible explanations for a model's decision-making. However, configuring an explainable model that effectively communicates among computational modules has received less attention. A recently proposed shared global workspace theory showed that networks of distributed modules can benefit from sharing information with a bottlenecked memory because the communication constraints encourage specialization, compositionality, and synchronization among the modules. Inspired by this, we propose Concept-Centric Transformers, a simple yet effective configuration of the shared global workspace for interpretability, consisting of: i) an object-centric-based memory module for extracting semantic concepts from input features, ii) a cross-attention mechanism between the learned concept and input embeddings, and iii) standard classification and explanation losses to allow human analysts to directly assess an explanation for the model's classification reasoning. We test our approach against other existing concept-based methods on classification tasks for various datasets, including CIFAR100, CUB-200-2011, and ImageNet, and we show that our model achieves better classification accuracy than all baselines across all problems but also generates more consistent concept-based explanations of classification output.

****

What's in the Flow? Exploiting Temporal Motion Cues for Unsupervised Generic Event Boundary Detection

Sourabh Vasant Gothe, Vibhav Agarwal, Sourav Ghosh, Jayesh Rajkumar Vachhani, Pranay Kashyap, Barath Raj Kandur Raja; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6941-6950

Generic Event Boundary Detection (GEBD) task aims to recognize generic, taxonomy-free boundaries that segment a video into meaningful events. Current methods typically involve a neural model trained on a large volume of data, demanding substantial computational power and storage space. We explore two pivotal questions pertaining to GEBD: Can non-parametric algorithms outperform unsupervised neural methods? Does motion information alone suffice for high performance? This inquiry drives us to algorithmically harness motion cues for identifying generic event boundaries in videos. In this work, we propose FlowGEBD, a non-parametric, unsupervised technique for GEBD. Our approach entails two algorithms utilizing optical flow: (i) Pixel Tracking and (ii) Flow Normalization. By conducting thorough experimentation on the challenging Kinetics-GEBD and TAPOS datasets, our results establish FlowGEBD as the new state-of-the-art (SOTA) among unsupervised methods. FlowGEBD exceeds the neural models on the Kinetics-GEBD dataset by obtaining an F1@0.05 score of 0.713 with an absolute gain of 31.7% compared to the unsupervised baseline and achieves an average F1 score of 0.623 on the TAPOS validation dataset.

****

SyntheWorld: A Large-Scale Synthetic Dataset for Land Cover Mapping and Building Change Detection

Jian Song, Hongruixuan Chen, Naoto Yokoya; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8287-8296

Synthetic datasets, recognized for their cost effectiveness, play a pivotal role in advancing computer vision tasks and techniques. However, when it comes to remote sensing image processing, the creation of synthetic datasets becomes challenging due to the demand for larger-scale and more diverse 3D models. This complexity is compounded by the difficulties associated with real remote sensing datasets, including limited data acquisition and high annotation costs, which amplifies the need for high-quality synthetic alternatives. To address this, we present SyntheWorld, a synthetic dataset unparalleled in quality, diversity, and scale. It includes 40,000 images with submeter-level pixels and fine-grained land cover annotations of eight categories, and it also provides 40,000 pairs of bitempor al image pairs with building change annotations for building change detection. We conduct experiments on multiple benchmark remote sensing datasets to verify the effectiveness of SyntheWorld and to investigate the conditions under which our synthetic data yield advantages. The dataset is available at https://github.com/JTRNEO/SyntheWorld.

****

Generalizing to Unseen Domains in Diabetic Retinopathy Classification

Chamuditha Jayanga Galappaththige, Gayal Kuruppu, Muhammad Haris Khan; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7685-7695

Diabetic retinopathy (DR) is caused by long-standing diabetes and is among the fifth leading cause for visual impairment. The prospects of early diagnosis and treatment could be helpful in curing the disease, however, the detection procedure is rather challenging and mostly tedious. Therefore, automated diabetic retinopathy classification using deep learning techniques has gained interest in the medical imaging community. Akin to several other real-world applications of deep learning, the typical assumption of i.i.d data is also violated in DR classification that relies on deep learning. Therefore, developing DR classification methods robust to unseen distributions is of great value. In this paper, we study the problem of generalizing a model to unseen distributions or domains (a.k.a domain generalization) in DR classification. To this end, we propose a simple and effective domain generalization (DG) approach that achieves self-distillation in vision transformers (ViT) via a novel prediction softening mechanism. This prediction softening is an adaptive convex combination of one-hot labels with the model's own knowledge. We perform extensive experiments on challenging open-source DR classification datasets under both multi-source and more challenging single-source DG settings with three different ViT backbones to establish the efficacy and applicability of our approach against competing methods. For the first time, we report the performance of several state-of-the-art domain generalization (DG) methods on open-source DR classification datasets after conducting thorough experiments. Finally, our method is also capable of delivering improved calibration performance than other methods, showing its suitability for safety-critical applications, including healthcare. We hope that our contributions would instigate more DG research across the medical imaging community.

****

A Generic and Flexible Regularization Framework for NeRFs

Thibaud Ehret, Roger Marí, Gabriele Facciolo; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 3088-3097

Neural radiance fields, or NeRF, represent a breakthrough in the field of novel view synthesis and 3D modeling of complex scenes from multi-view image collections. Numerous recent works have shown the importance of making NeRF models more robust, by means of regularization, in order to train with possibly inconsistent and/or very sparse data. In this work, we explore how differential geometry can provide elegant regularization tools for robustly training NeRF-like models, which are modified so as to represent continuous and infinitely differentiable functions. In particular, we present a generic framework for regularizing different types of NeRFs observations to improve the performance in challenging conditions. We also show how the same formalism can also be used to natively encourage the regularity of surfaces by means of Gaussian or mean curvatures.

****

MarsLS-Net: Martian Landslides Segmentation Network and Benchmark Dataset

Sidike Paheding, Abel A. Reyes, A. Rajaneesh, K.S. Sajinkumar, Thomas Oommen; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8236-8245

Martian landslide segmentation is a challenging task compared to the same task on Earth. One of the reasons is that vegetation is typically lost or significantly less compared to its surroundings in the regions of landslide on Earth. In contrast, Mars is a desert planet, and there is no vegetation to aid landslide detection and segmentation. Recent work has demonstrated the strength of vision transformer (ViT) based deep learning models for various computer vision tasks. Insp ired by the multi-head attention mechanism in ViT, which can model the global long-range spatial correlation between local regions in the input image, we hypothesize self-attention mechanism can effectively capture pertinent contextual information for the Martian landslide segmentation task. Furthermore, considering parameter efficiency or model size is another important factor for deep learning algorithms, we construct a new feature representation block, namely Progressively Expanded Neuron Attention (PEN-Attention), to extract more relevant features with significantly fewer trainable parameters. Overall, we refer to our deep learning architecture as the Martian landslide segmentation network (MarsLS-Net). In addition to the new architecture, we introduce a new multi-modal Martian landslide segmentation dataset for the first time, which will be made publicly available at https://github.com/MAIN-Lab/Multimodal-Martian-Landslides-Dataset

****

You Can Run but Not Hide: Improving Gait Recognition With Intrinsic Occlusion Type Awareness

Ayush Gupta, Rama Chellappa; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5893-5902

While gait recognition has seen many advances in recent years, the occlusion problem has largely been ignored. This problem is especially important for gait recognition from uncontrolled outdoor sequences at range - since any small obstruction can affect the recognition system. Most current methods assume the availability of complete body information while extracting the gait features. When parts of the body are occluded, these methods may hallucinate and output a corrupted gait signature as they try to look for body parts which are not present in the input at all. To address this, we exploit the learned occlusion type while extracting identity features from videos. Thus, in this work, we propose an occlusion aware gait recognition method which can be used to model intrinsic occlusion awareness into potentially any state-of-the-art gait recognition method. Our experiments on the challenging GREW and BRIAR datasets show that networks enhanced with this occlusion awareness perform better at recognition tasks than their counterparts trained on similar occlusions.

****

SphereCraft: A Dataset for Spherical Keypoint Detection, Matching and Camera Pose Estimation

Christiano Gava, Yunmin Cho, Federico Raue, Sebastian Palacio, Alain Pagani, Andreas Dengel; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4408-4417

This paper introduces SphereCraft, a dataset specifically designed for spherical keypoint detection, matching, and camera pose estimation. The dataset addresses the limitations of existing datasets by providing extracted keypoints from various detectors, along with their ground truth correspondences. Synthetic scenes with photo-realistic rendering and accurate 3D meshes are included, as well as real-world scenes acquired from different spherical cameras. SphereCraft enables the development and evaluation of algorithms targeting multiple camera viewpoints, advancing the state-of-the-art in computer vision tasks involving spherical images. Our dataset is available at https://dfki.github.io/spherecraftweb/.

****

Best of Both Worlds: Learning Arbitrary-Scale Blind Super-Resolution via Dual Degradation Representations and Cycle-Consistency

Shao-Yu Weng, Hsuan Yuan, Yu-Syuan Xu, Ching-Chun Huang, Wei-Chen Chiu; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1547-1556

Single image super-resolution (SISR) for reconstructing from a low-resolution (LR) input image its corresponding high-resolution (HR) output is a widely-studied research problem in the field of multimedia applications and computer vision. Despite the magic leap brought by recent development of deep neural networks for SISR, such problem is still considered to be quite challenging and non-scalable for the real-world data due to its ill-posed nature, where the degradations happened to the input LR images are usually complex and even unknown (in which the degradations in the test data could be unseen or different from the ones shown in the training dataset). To this end, two branches of SISR methods have emerged: blind super-resolution (blind-SR) and arbitrary-scale super-resolution (ASSR), where the former aims to reconstruct SR images under the unknown degradations, while the latter improves the scalability via learning to handle arbitrary up-sampling ratios. In this paper, we propose a holistic framework to take both blind-SR and ASSR tasks (accordingly named as arbitrary-scale blind-SR) into consideration with two main designs: 1) learning dual degradation representations where the implicit and explicit representations of degradation are sequentially extracted from the input LR image, and 2) modeling both upsampling (i.e. LR to HR) and downsampling (i.e. HR to LR) processes at the same time, where they utilize the implicit and explicit degradation representations respectively, in order to enable the cycle-consistency objective and further improve the training. We conduct extensive experiments on various datasets where the results well verify the effectiveness of our proposed framework in handling complex degradations as well as its superiority with respect to several state-of-the-art baselines.

****

VMFormer: End-to-End Video Matting With Transformer

Jiachen Li, Vidit Goel, Marianna Ohanyan, Shant Navasardyan, Yunchao Wei, Humphrey Shi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6678-6687

Video matting aims to predict the alpha mattes for each frame from a given input video sequence. Recent solutions to video matting have been dominated by deep convolutional neural networks (CNN) for the past few years, which have become the de-facto standard for academia and industry. However, they have the inbuilt inductive bias of locality and do not capture the global characteristics of an image due to the CNN-based architectures. They also need long-range temporal modeling considering computational costs when dealing with feature maps of multiple frames. In this paper, we propose VMFormer: a transformer-based end-to-end method for video matting. It makes predictions on alpha mattes of each frame from learnable queries given a video input sequence. Specifically, it leverages self-attention layers to build global integration of feature sequences with short-range temporal modeling on successive frames. We further apply queries to learn global representations through cross-attention in the transformer decoder with long-range temporal modeling upon all queries. In the prediction stage, both queries and corresponding feature maps are used to make the final prediction of alpha mattes. Experiments show that VMFormer outperforms previous CNN-based video matting methods on synthetic benchmarks with different input resolutions, as an end-to-end video matting solution built upon a full vision transformer with predictions on the learnable queries. The project is open-sourced at https://chrisjuniorli.github.io/project/VMFormer.

****

Feed-Forward Latent Domain Adaptation

Ondrej Bohdal, Da Li, Shell Xu Hu, Timothy Hospedales; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8490-8499

We study a new highly-practical problem setting that enables resource-constrained edge devices to adapt a pre-trained model to their local data distributions. Recognizing that device's data are likely to come from multiple latent domains that include a mixture of unlabelled domain-relevant and domain-irrelevant examples, we focus on the comparatively under-studied problem of latent domain adaptation. Considering limitations of edge devices, we aim to only use a pre-trained model and adapt it in a feed-forward way, without using back-propagation and without access to the source data. Modelling these realistic constraints bring us to the novel and practically important problem setting of feed-forward latent domain adaptation. Our solution is to meta-learn a network capable of embedding the mixed-relevance target dataset and dynamically adapting inference for target examples using cross-attention. The resulting framework leads to consistent improvements over strong ERM baselines. We also show that our framework sometimes even improves on the upper bound of domain-supervised adaptation, where only domain-relevant instances are provided for adaptation. This suggests that human annotated domain labels may not always be optimal, and raises the possibility of doing better through automated instance selection.

****

Triplet Attention Transformer for Spatiotemporal Predictive Learning

Xuesong Nie, Xi Chen, Haoyuan Jin, Zhihang Zhu, Yunfeng Yan, Donglian Qi; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 7036-7045

Spatiotemporal predictive learning offers a self-supervised learning paradigm that enables models to learn both spatial and temporal patterns by predicting future sequences based on historical sequences. Mainstream methods are dominated by recurrent units, yet they are limited by their lack of parallelization and often underperform in real-world scenarios. To improve prediction quality while maintaining computational efficiency, we propose an innovative triplet attention transformer designed to capture both inter-frame dynamics and intra-frame static features. Specifically, the model incorporates the Triplet Attention Module (TAM), which replaces traditional recurrent units by exploring self-attention mechanisms in temporal, spatial, and channel dimensions. In this configuration: (i) temporal tokens contain abstract representations of inter-frame, facilitating the capture of inherent temporal dependencies; (ii) spatial and channel attention combine to refine the intra-frame representation by performing fine-grained interactions across spatial and channel dimensions. Alternating temporal, spatial, and channel-level attention allows our approach to learn more complex short- and long-range spatiotemporal dependencies. Extensive experiments demonstrate performance surpassing existing recurrent-based and recurrent-free methods, achieving state-of-the-art under multi-scenario examination including moving object trajectory prediction, traffic flow prediction, driving scene prediction, and human motion capture.

****

Arbitrary-Resolution and Arbitrary-Scale Face Super-Resolution With Implicit Representation Networks

Yi Ting Tsai, Yu Wei Chen, Hong-Han Shuai, Ching-Chun Huang; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 4270-4279

Face super-resolution (FSR) is a critical technique for enhancing low-resolution facial images and has significant implications for face-related tasks. However, existing FSR methods are limited by fixed up-sampling scales and sensitivity to input size variations. To address these limitations, this paper introduces an Arbitrary-Resolution and Arbitrary-Scale FSR method with implicit representation networks (ARASFSR), featuring three novel designs. First, ARASFSR employs 2D deep features, local relative coordinates, and up-sampling scale ratios to predict RGB values for each target pixel, allowing super-resolution at any up-sampling scale. Second, a local frequency estimation module captures high-frequency facial texture information to reduce the spectral bias effect. Lastly, a global coordinate modulation module guides FSR to leverage prior knowledge of facial structure effectively and achieve resolution adaptation. Quantitative and qualitative evaluations demonstrate the robustness of ARASFSR over existing state-of-the-art methods while super-resolving facial images across various input sizes and up-sampling scales.

****

FPGAN-Control: A Controllable Fingerprint Generator for Training With Synthetic Data

Alon Shoshan, Nadav Bhonker, Emanuel Ben Baruch, Ori Nizan, Igor Kviatkovsky, Joshua Engelsma, Manoj Aggarwal, Gérard Medioni; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 6067-6076

Training fingerprint recognition models using synthetic data has recently gained increased attention in the biometric community as it alleviates the dependency on sensitive personal data. Existing approaches for fingerprint generation are limited in their ability to generate diverse impressions of the same finger, a key property for providing effective data for training recognition models. To address this gap, we present FPGAN-Control, an identity preserving image generation

framework which enables control over the fingerprint's image appearance (e.g., fingerprint type, acquisition device, pressure level) of generated fingerprints. We introduce a novel appearance loss that encourages disentanglement between the fingerprint's identity and appearance properties. In our experiments, we used the publicly available NIST SD302 (N2N) dataset for training the FPGAN-Control model. We demonstrate the merits of FPGAN-Control, both quantitatively and qualitatively, in terms of identity preservation level, degree of appearance control, and low synthetic-to-real domain gap. Finally, training recognition models using only synthetic datasets generated by FPGAN-Control lead to recognition accuracies that are on par or even surpass models trained using real data. To the best of our knowledge, this is the first work to demonstrate this.
****

Continual Learning of Unsupervised Monocular Depth From Videos

Hemang Chawla, Arnav Varma, Elahe Arani, Bahram Zonooz; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 8419-8429

Spatial scene understanding, including monocular depth estimation, is an important problem with various applications such as robotics and autonomous driving. While improvements in unsupervised monocular depth estimation have potentially allowed models to be trained on diverse crowdsourced videos, this remains under-explored as most methods utilize the standard training protocol wherein the models are trained from scratch on all data after new data is collected. Instead, continual training of models on sequentially collected data would significantly reduce computational and memory costs. Nevertheless, naive continual training leads to catastrophic forgetting, where the model performance deteriorates on older domains as it learns on newer domains, highlighting the trade-off between model stability and plasticity. While several techniques have been proposed to address this issue in image classification, the high-dimensional and spatiotemporally correlated outputs of depth estimation make it a distinct challenge. To the best of our knowledge, no framework or method currently exists focusing on the problem of continual learning in depth estimation. Thus, we introduce a framework that captures the challenges of continual unsupervised depth estimation (CUDE), and define the necessary metrics for evaluating model performance. We propose a rehearsal-based dual-memory method MonoDepthCL, which utilizes spatiotemporal consistency for continual learning in depth estimation, even when the camera intrinsics are unknown.
****

CXR-IRGen: An Integrated Vision and Language Model for the Generation of Clinically Accurate Chest X-Ray Image-Report Pairs

Junjie Shentu, Noura Al Moubayed; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 5212-5221

Chest X-Ray (CXR) images play a crucial role in clinical practice, providing vital support for diagnosis and treatment. Augmenting the CXR dataset with synthetically generated CXR images annotated with radiology reports can enhance the performance of deep learning models for various tasks. However, existing studies have primarily focused on generating unimodal data of either images or reports. In this study, we propose an integrated model, CXR-IRGen, designed specifically for generating CXR image-report pairs. Our model follows a modularized structure consisting of a vision module and a language module. Notably, we present a novel prompt design for the vision module by combining both text embedding and image embedding of a reference image. Additionally, we propose a new CXR report generation model as the language module, which effectively leverages a large language model and self-supervised learning strategy. Experimental results demonstrate that our new prompt is capable of improving the general quality (FID) and clinical efficacy (AUROC) of the generated images, with average improvements of 15.84% and 1.84%, respectively. Moreover, the proposed CXR report generation model outperforms baseline models in terms of clinical efficacy (F1 score) and exhibits a high-level alignment of image and text, as the best F1 score of our model is 6.93% higher than the state-of-the-art CXR report generation model. Our code is available at https://github.com/junjie-shentu/CXR-IRGen.

****

Overcoming Catastrophic Forgetting for Multi-Label Class-Incremental Learning

Xiang Song, Kuang Shu, Songlin Dong, Jie Cheng, Xing Wei, Yihong Gong; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 2389-2398

Despite the recent progress of class-incremental learning (CIL) methods, their capabilities in real-world scenarios such as multi-label settings remain unexplored. This paper focuses on a more practical CIL problem named multi-label class-incremental learning (MLCIL). MLCIL requires the vision models to overcome catastrophic forgetting of old knowledge while learning new classes from multi-label samples. Direct application of existing CIL methods to MLCIL leads to label absence, representative sample selection, and feature dilution problems. To address these problems, we present a novel AdaPtive Pseudo-Label-drivEn (APPLE) framework consisting of three components. First, the adaptive pseudo-label strategy is proposed to solve the label absence problem, which leverages the old model to annotate old classes for new samples. Second, a cluster sampling strategy is proposed to obtain more diverse samples to alleviate catastrophic forgetting under the MLCIL setting better. Finally, a class attention decoder is designed to mitigate the object feature dilution problem in multi-label samples. The extensive experiments on PASCAL VOC 2007 and MS-COCO demonstrate that our proposed method significantly outperforms other representative state-of-the-art CIL methods.

****

PatchRefineNet: Improving Binary Segmentation by Incorporating Signals From Optimal Patch-Wise Binarization

Savinay Nagendra, Daniel Kifer; Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2024, pp. 1361-1372

The purpose of binary segmentation models is to determine which pixels belong to an object of interest (e.g., which pixels in an image are part of roads). The models assign a logit score (i.e., probability) to each pixel and these are converted into predictions by thresholding (i.e., each pixel with logit score >= t is predicted to be part of a road). However, a common phenomenon in current and former state-of-the-art segmentation models is spatial bias -- in some patches, the logit scores are consistently biased upwards and in others they are consistently biased downwards. These biases cause false positives and false negatives in the final predictions. In this paper, we propose PatchRefineNet (PRN), a small network that sits on top of a base segmentation model and learns to correct its patch-specific biases. Across a wide variety of base models, PRN consistently helps them improve mIoU by 2-3%. One of the key ideas behind PRN is the addition of a novel supervision signal during training. Given the logit scores produced by the base segmentation model, each pixel is given a pseudo-label that is obtained by optimally thresholding the logit scores in each image patch. Incorporating these pseudo-labels into the loss function of PRN helps correct systematic biases and reduce false positives/negatives. Although we mainly focus on binary segmentation, we also show how PRN can be extended to saliency detection and few-shot segmentation. We also discuss how the ideas can be extended to multi-class segmentation. Source code is available at https://github.com/savinay95n/PatchRefineNet.

****