Project List

Chi Keung Tang

NeRF-RPN: A general framework for object detection in NeRF This paper presents the first significant object detection framework, NeRF-RPN, which directly operates on NeRF. Given a pre-trained NeRF model, NeRF-RPN aims to detect all bounding boxes of objects in a scene. By exploiting a novel voxel representation that incorporates multi-scale 3D neural volumetric features, we demonstrate it is possible to regress the 3D bounding boxes of objects in NeRF directly without rendering the NeRF at any viewpoint. NeRF-RPN is a general framework and can be applied to detect objects without class labels. We experimented the NeRF-RPN with various backbone architectures, RPN head designs and loss functions. All of them can be trained in an end-to-end manner to estimate high quality 3D bounding boxes. To facilitate future research in object detection for NeRF, we built a new benchmark dataset which consists of both synthetic and real-world data with careful labeling and clean up. Code and dataset will be made available. Demo link: NeRF-RPN

FLNeRF: 3D Facial Landmarks Estimation in Neural Radiance Fields This paper presents the first significant work on directly predicting 3D face landmarks on neural radiance fields (NeRFs), without using any intermediate representations such as 2D images, depth maps, or point clouds. Our 3D coarse-to-fine Face Landmarks NeRF (FLNeRF) model efficiently samples from the NeRF on the whole face with individual facial features for accurate landmarks. To mitigate the limited number of facial expressions in the available data, local and non-linear NeRF warp is applied at facial features in fine scale to simulate large emotions range, including exaggerated facial expressions (e.g., cheek blowing, wide opening mouth, eye blinking), for training FLNeRF. With such expression augmentation, our model can predict 3D landmarks not limited to the 20 discrete expressions given in the data. Robust 3D NeRF facial landmarks contribute to many downstream tasks. As an example, we modify MoFaNeRF to enable high-quality face editing and swapping using face landmarks on NeRF, allowing more direct control and wider range of complex expressions. Experiments show that the improved model using landmarks achieves comparable to better results. Demo link: FLNeRF

Continual Test-Time Generalizable Domain Adaptation for Robust Object Detection Real-world environment can be highly dynamic causing substantial domain shifts. Such real-world domain shifts can span over time with domain changes across multiple domains, manifested into the pertinent content or style changes, or both. Performance of safety-critical applications, especially robust object detection system in autonomous driving, must adapt to such test-time domain shifts. However, our empirical analysis shows existing domain adaptation and generalization methods fail to fit the domain changes with substantial style or content shifts. To simultaneously address temporal and multiple domain shifts, we propose generalizable domain adaptation method in test time for object detection, which consists of the following three collaborative modules: 1) Domain generalization training (DGT) module initializes a style-invariant object detection model; 2) Test-time adaptation (TTA) module updates the DGT trained model online during inference; 3) Generalizable weights preservation (GWP) module keeps the learned generalizable weights to avoid domain overfitting in generalization across multiple domains. Extensive experiments demonstrate these three modules collaboratively enable a deep model to generalize well under challenging real-world domain shifts.

Mask-Free Video Instance Segmentation The recent advancements in Video Instance Segmentation (VIS) has largely been driven by the use of deeper and increasingly data-hungry transformer-based models. However, video masks labels are tedious and expensive to annotate, limiting the scale and diversity of existing VIS datasets. In this work, we aim to remove the mask-annotation requirement. We propose MaskFreeVIS, achieving highly competitive VIS performance, while only using bounding box annotations for the object state. We leverage the rich temporal mask consistency constraints in videos by introducing the Temporal KNN-patch Loss (TK-Loss), providing strong mask supervision without any labels. Our TK-Loss finds one-to-many matches across frames, through an efficient patch-matching step followed by a K-nearest neighbor selection. A consistency loss is then enforced on the found matches. Our mask-free objective is simple to implement, has no trainable parameters, is computationally efficient, yet outperforms baselines employing, e.g., state-of-the-art optical flow to enforce temporal mask consistency. We validate MaskFreeVIS on the YouTube-VIS 2019/2021, OVIS and BDD100K MOTS benchmarks. The results clearly demonstrate the efficacy of our method by drastically narrowing the gap between fully and weakly-supervised VIS performance. Our codes and trained models will be made publicly available.

A New Basic Framework for Egocentric Hand-Object Interaction Understanding Egocentric Hand-Object Interaction (Ego-HOI) has drawn huge attention recently. Some large-scale datasets such as Ego4D and EPIC-KITCHENS have been proposed to promote Ego-HOI. However, most of the current works still rely on existing tools and paradigms from third-person video action recognition. Due to the large domain gap and different properties inherent in egocentric and exocentric action videos, adopting the settings of third-person action understanding is arguably suboptimal. In this work, we first put forward a new framework to advance Ego-HOI recognition by baseline design, comprehensive pre-train set reorganization, balanced test set construction, and training-finetuning strategy. With our new framework, we achieve not only state-of-the-art performance on Ego-HOI benchmarks but also several new and effective mechanisms and settings to advance further research. We believe our data and the findings will pave a new way for Ego-HOI understanding. Our code and data will be publicly available.

Fuse-HOI: Advancing Human-Object Interaction Detection by Exploiting Non-HOI Data Human-Object Interaction (HOI) detection plays an important role in human activity understanding. Costly annotation of HOI data, especially for rare HOI where sufficient data are lacking, has hindered research progress and performance. In this paper, instead of manually labeling HOI data, we leverage non-HOI datasets with human action labels (human box and action class), which are rich in semantics and widely used in other tasks (e.g., video action recognition). We conduct thorough experiments to analyze how human pose, human body part, and natural language in non-HOI data can effectively supervise the learning of HOI models, and conclude empirically that performance can be significantly improved by incorporating the combination of these components. Based on the analysis, we propose a multi-modal training paradigm that can be adapted to various HOI detectors during training. Note that although we use additional features for training, following our training paradigm, these features are not required in testing. Since only HOI detector is retained during inference, no extra computation cost is involved. Our paradigm demonstrates high effectiveness and flexibility. Compared with the state-of-the-arts, the experimental results on HICO-DET and V-COCO demonstrate significant performance advancement of our method.

Ultrahigh Resolution Image/Video Matting with Spatio-Temporal Sparsity Commodity ultra-high definition (UHD) displays are becoming more affordable which demand imaging in ultra high resolution (UHR). This paper proposes SparseMat, a computationally efficient approach for UHR image/video matting. Note that it is infeasible to directly process UHR images at full resolution in one shot using existing matting algorithms without running out of memory on consumer-level computational platforms, e.g., Nvidia 1080Ti with 11G memory, while patch-based approaches can introduce unsightly artifacts due to patch partitioning. Instead, our method resorts to spatial and temporal sparsity for solving general UHR matting. During processing videos, huge computation redundancy can be reduced through the rational use of spatial and temporal sparsity. In this paper, we show how to effectively estimate spatio-temporal sparsity, which serves as a gate to activate input pixels for the matting model. Under the guidance of such sparsity, our method discards patch-based inference in lieu of memory-efficient and full-resolution matte refinement. Extensive experiments demonstrate that SparseMat can effectively and efficiently generate high-quality alpha matte for UHR images and videos in one shot. Codes will be made available.

H-VFI: Hierarchical Frame Interpolation for Videos with Large Motions Capitalizing on the rapid development of neural networks, recent video frame interpolation (VFI) methods have achieved notable improvements. However, they still fall short for real-world videos containing large motions. Complex deformation and/or occlusion caused by large motions make it an extremely difficult problem in video frame interpolation. In this paper, we propose a simple yet effective solution, H-VFI, to deal with large motions in video frame interpolation. H-VFI contributes a hierarchical video interpolation transformer (HVIT) to learn a deformable kernel in a coarse-to-fine strategy in multiple scales. The learnt deformable kernel is then utilized in convolving the input frames for predicting the interpolated frame. Starting from the smallest scale, H-VFI updates the deformable kernel by a residual in succession based on former predicted kernels, intermediate interpolated results and hierarchical features from transformer. Bias and masks to refine the final outputs are then predicted by a transformer block based on interpolated results. The advantage of such a progressive approximation is that the large motion frame interpolation problem can be decomposed into several relatively simpler sub-tasks, which enables a very accurate prediction in the final results. Another noteworthy contribution of our paper consists of a large-scale high-quality dataset, YouTube200K, which contains videos depicting a great variety of scenarios captured at high resolution and high frame rate. Extensive experiments on multiple frame interpolation benchmarks validate that H-VFI outperforms existing state-of-the-art methods especially for videos with large motions.