WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wild

Rolandos Alexandros Potamias, Jinglei Zhang,
Jiankang Deng, Stefanos Zafeiriou

Paper

This paper introduces a new pipeline for efficient 3D multi-hand reconstruction from single-view images in real-world, uncontrolled environments. The pipeline includes real-time fully convolutional hand localization and high-fidelity transformer-based 3D hand reconstruction components. Additionally, they introduce a 2 million hand image dataset.

Motivation

Previous methods for 3D hand pose estimation focused on images containing a fixed number of hands and cannot generalize to in-the-wild images. Also, there was a lack of large-scale, annotated datasets for real-world hand detection and 3D pose estimation.

Architecture

Hand Detection and Localization

The DarkNet model is used as the backbone to generate feature maps from the image. The last three feature maps are then used to generate a multi-scale feature pyramid, with the Path Aggregation Network, which is an extension of the Feature Pyramid network. The features are passed through three detection heads to determine bounding boxes and if the hand is left or right.

Hand Reconstruction

The image is first split into patches, and then embedded to high dimensional tokens. Positional embeddings are added to the image tokens to uniquely encode their spatial location. These tokens are passed through a Vision Transformer (ViT) encoder backbone along with learnable pose, shape, and camera tokens to generate the new image, pose, shape, and camera feature tokens. These feature tokens are regressed to a rough estimation of pose and shape MANO parameters.

The hand estimation is projected to a feature map using the estimated camera parameters. A set of deconvolution layers are used to upsample the feature map to multiple higher resolution feature maps. These feature maps are then used to refine the hand pose and mesh.

Limitations

1. Examples do not demonstrate the accuracy of hidden parts of the hand under severe occlusion.
2. Video frames are processed independently, and the model does not leverage temporal consistency for improved predictions for videos.