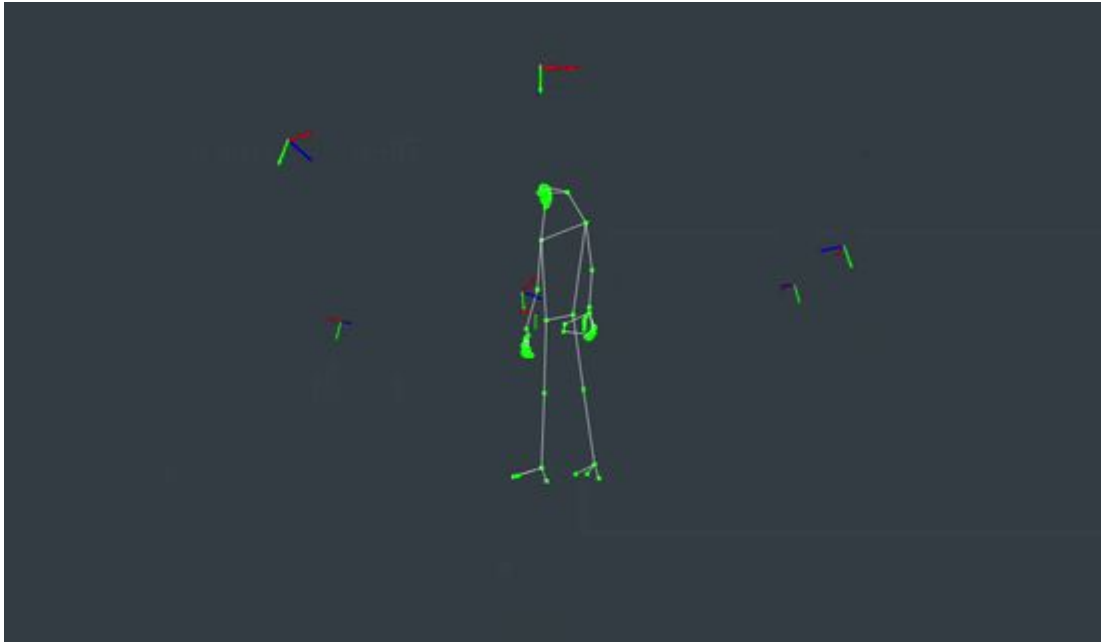
Motivation

- Hand details are not clearly visible from the Kinect cameras due to the distance
- The 360 camera has a closer view with limited occlusion
- Use the 360 camera to improve hand details



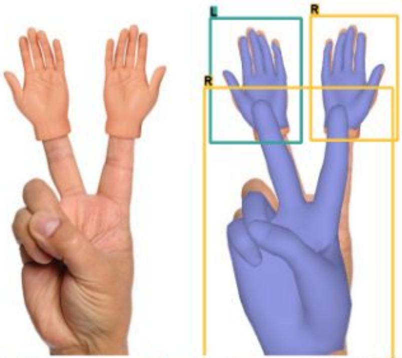
WiLoR: End-to-end 3D Hand Localization and Reconstruction in-the-wil

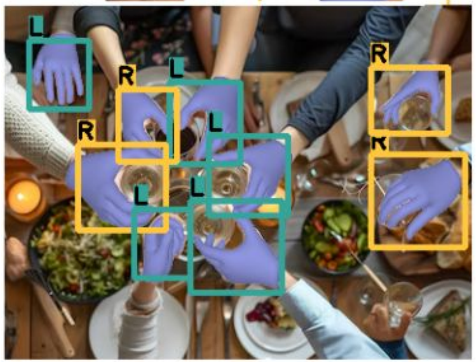
CVPR25

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

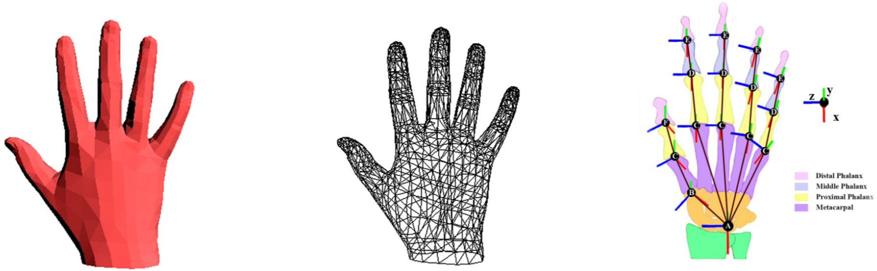
Imperial College London, Shanghai Jiao Tong University

Samples





MANO



(a) (b) (c)

Part of SMPL-X whole-body model

Parameters:

- Pose: 45 parameters (15 joints × 3 rotation parameters)
- Shape: 10 parameters
- Vertices: 778

Dataset

- 1,400 YouTube videos
- hand activities including sign language, cooking, everyday activities, sports, and games with ego- and exo-centric viewpoints

Dataset Annotation

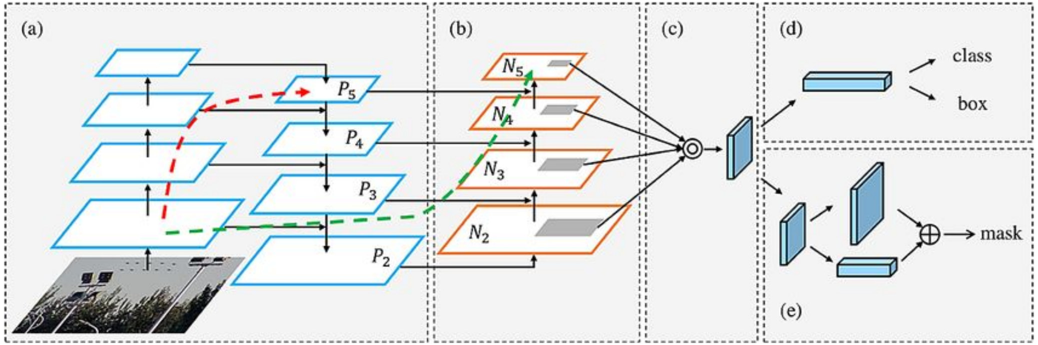
1. Hand Detection
    a. ViTPose
    b. AlphaPose
2. Hand Pose Estimation
    a. MediaPipe
    b. OpenPose
    c. ContactHands
3. Fine-tuning
    a. Confidence-based weighted average for hand localization
    b. 2D landmarks for 3D parametric hand model fitting
    c. Bio-mechanical constraints for rotations and bone length

$$\hat{y} = \frac{\sum_i P(\mathbf{b}_i|d_i)\mathbf{b}_i}{\sum_i P(\mathbf{b}_i|d_i)}$$

$$\mathcal{L}_{proj} = ||\mathbf{J}_{\mathcal{M}} - \pi(\hat{\mathbf{J}}_{\mathbf{s}}, K)||_1,$$

$$\mathcal{L}_{BMC} = \mathcal{L}_{BL} + \mathcal{L}_A$$

Detector



- BCE: Binary Cross Entropy
- DFL: Distributional Focal Loss
- IoU: Intersection over Union
- Kpts: L2 loss on the keypoints

$$\mathcal{L} = \lambda_0 \mathcal{L}_{BCE} + \lambda_1 \mathcal{L}_{DFL} + \lambda_2 \mathcal{L}_{CIoU} + \lambda_3 \mathcal{L}_{kpts}$$

3D Reconstruction

- Hand pose θ (48p)
- Hand shape β (10p)
- Camera parameters (translation and scale)

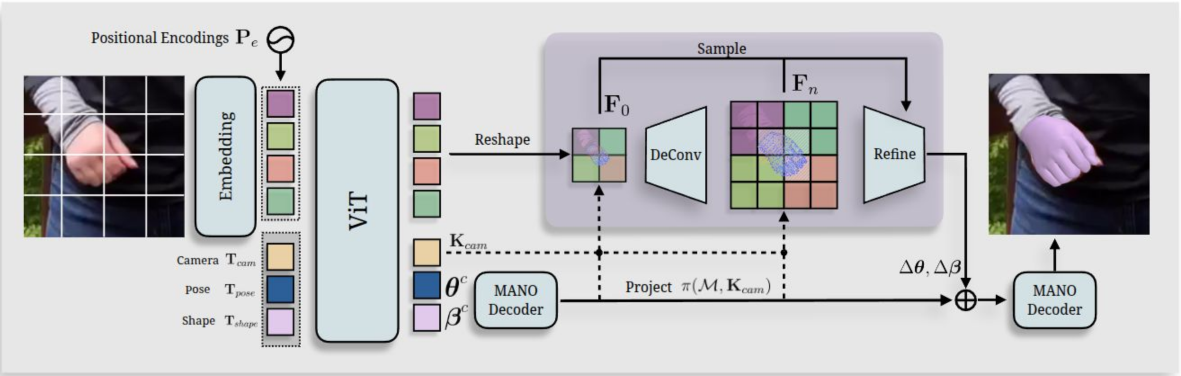
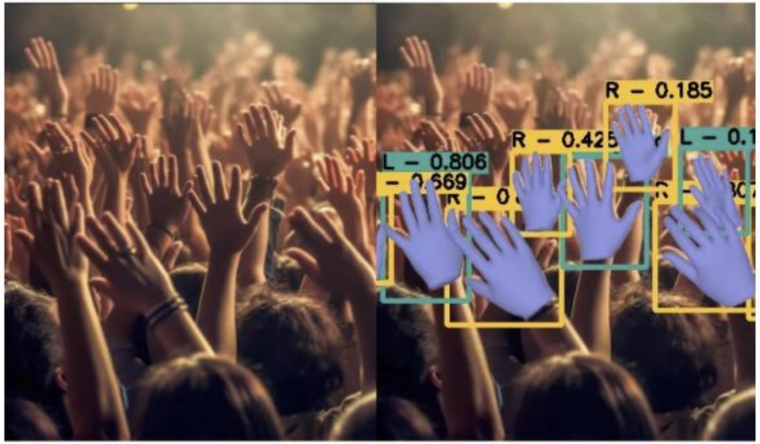


Figure 4. **Overview of the proposed 3D hand pose estimation method**: Given an image $\mathbf{I}_h$ represented as a series of feature tokens $\mathbf{T}_{img}$ along with a set of learnable camera $\mathbf{T}_{cam}$, pose $\mathbf{T}_{pose}$ and shape $\mathbf{T}_{shape}$ tokens, we initially predict a rough estimation of the MANO [74] and camera $\mathbf{K}_{cam}$ parameters using a ViT backbone (light blue). The updated image tokens are then reshaped and upsampled through a series of deconvolutional layers to form a set of multi-resolution feature maps $\{\mathbf{F}_0, ..., \mathbf{F}_0\}$. We then project the estimated 3D hand to the generated feature maps and sample image-aligned multi-scale features through a novel refinement module (purple). The sampled features are used to predict pose and shape residuals $\Delta\theta, \Delta\beta$ that refine the coarse hand estimation. Using this coarse-to-fine pose estimation strategy we facilitate image alignment and achieve better reconstruction performance.

Limitations

Learning-based models fail on edge cases



Detector might fail on small hands

Question?

Thanks!