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

Supplementary Material

1. Implementation Details

In this section we report the training details of the hand detection and the hand pose estimation models.

1.1. Hand Detection and Localization

To train the detector model we use WHIM dataset that comprises of over 2M in the wild images from daily activities. We train WiLoR detector with Adam optimizer for 200 epochs with early stopping if there is no loss decrease for over 30 epochs. Initiate the training with a learning rate of 0.01 and linearly decrease to $1e-6$ for the last 30 epochs of the training. We trained the model for three weeks using two NVIDIA RTX 4090 and a batch size of 256. To weight the different losses we set $\lambda_0 = 0.5$ for the classification loss, $\lambda_1 = 1.5$ for the distribution focal length loss, $\lambda_2 = 15$ for the bounding box loss, $\lambda_3 = 10$ for the keypoints loss. We use random mosaic augmentations with probability 0.7, random rotations between $[-60^o, 60^o]$ and random image scaling between [0.5, 1].

1.2. Hand Pose Estimation

We build our hand pose estimation method on top of a ViT-Large backbone with pre-trained weights from ViTPose [18], with a hidden dimension of 1280. Apart from the image patches, we use three additional learnable tokens that correspond to hand pose, shape and camera translation and scale. We initialize the tokens with the mean pose, shape and camera parameters from the training set. Using a set of fully-connected layers, we map the output tokens to 96 MANO pose parameters (15 joint rotations + 1 global orientation represented in 6d rotation format [21]), 10 MANO shape parameters and a 3D camera translation. We then reshape the output image tokens to a 16×12 image form, and perform two sets of upsamplings using deconvolutions. At each upsampling step we reduce the feature by 2 times. Using the initially estimated camera parameters we project the rough MANO estimation to the feature maps and sample a set of multi-scale per-vertex image-aligned features. The concatenated set of features is then aggregated and regressed from a set of fully-connected layers that predict the pose, shape and camera residuals. We train the model for 1000 epochs using Adam optimizer with an initial learning rate of 1e-5 and a weight decay of 1e-4. Similar to the hand detector, we apply random scaling, rotations and color jitter during training. Similar to [13], to balance the losses we set $\lambda_{3D} = 0.05$, $\lambda_{2D} = 0.01$, $\lambda_{pose} = 0.001$, $\lambda_{shape} = 0.0005$ and $\lambda_{adv} = 0.0005$.

2. Comparison with existing datasets

In contrast to 3D hand pose estimation methods that utilizes images of tightly cropped hands, to train a powerful hand detector network, it is required to create a dataset that contains images with multiple hands under different occlusions, views, illuminations and skin tones. Bellow we compare WHIM with such available datasets. WHIM is 100× larger than previous in-the-wild multi-hand datasets.

Dataset	#Img	Annotations	Egocentric	Third-Person	Objects	Real	3D
OxfordHands	13K	Manual	✘	✔	✘	✔	✘
MPI-HP	25K	Manual	✘	✔	✘	✔	✘
Coco-Whole	200K	Manual	✘	✔	✘	✔	✘
BEDLAM	380K	GT	✘	✔	✘	✘	✔
AGORA	18K	GT	✘	✔	✘	✘	✔
ContactHands	21K	Manual	✘	✔	✔	✔	✘
CocoHands	25K	Manual	✘	✔	✔	✔	✘
BodyHands	20K	Manual	✘	✔	✘	✔	✘
WHIM	**2M**	Auto	✔	✔	✔	✔	✔

Table 1. **Comparison with existing hand datasets.** WHIM is 100× larger than previous multi-hand dataset.s

3. Limitations

Although achieving state-of-the-art performance on both 3D hand pose estimation and hand detection tasks, WiLoR still fails to recover challenging cases. Despite being trained on a large-scale dataset, the data distribution is still limited to 'common' hand poses and appearances, failing to generalize to samples far from the trained distribution. As can be seen in Fig. 1 WiLoR can fail under extreme finger poses and can also fail to detect hands in crowded environments. Creating a synthetic dataset with diverse hand poses and photorealistic hands could help mitigate these issues [14]. Additionally, since WiLoR employs a bottom-up reconstruction strategy, interactions and contacts between hands may not be adequately captured in 3D space. In scenarios where accurate hand contact estimation is crucial [1, 24], incorporating additional interaction constraints [17] may be necessary. Finally, WiLoR estimates 3D hand poses in camera space, which may lead to inaccurate assumptions about the overall 3D scene. Adapting WiLoR with a 3D metric foundational model [19, 20] could enable more accurate 3D reconstruction in world space.

4. Training Datasets

To train our hand pose estimation module we use a combination of datasets to enforce the generalization of the model. In particular, we use 14 datasets with both 2D and 3D annotations, from three major categories: controlled environ-

Figure 1. **Failure Cases**. WiLoR can still fail to reconstruct complex finger poses or detect small hands in crowded environment.

ment hand images, hand-object interaction, in-the-wild and synthetic datasets, resulting in 4.2M images total:

- FreiHAND [23] is a common 3D hand pose estimation dataset composed of 132K images of indoor, outdoor, and synthetic scenes. It provides both 3D hand and 2D keypoint annotations.
- MTC [16] is a subset of Panoptic Studio Dataset [10] that contains 360K multi-view images in a studio environment. The dataset provides both 3D hand and 2D keypoint annotations.
- InterHand2.6M [11] is a large scale environment from light stage environment that contains hand articulations from 27 different subjects and 80 different cameras. The dataset provides both 3D hand and 2D keypoint annotations.

To increase the generalization of WiLoR under severe occlusions we include several datasets where hands interact with objects.

- HO3D [8] provides a hand-object dataset with over than 120K images from multi-view cameras of hands interacting with objects. It is used as one of the main benchmarks for hand and object reconstruction. Images were captured in an lab environment setting. It provides both 3D hand and 2D keypoint annotations.
- H2O3D [8] contains over 60K images from five multi-view cameras of hands interacting with objects. In contrast to HO3D, each subject is interacting with the object using two hands which increases the occlusions of the hands. It provides both 3D hand and 2D keypoint annotations.
- DEX YCB [5] similar to HO3D, DEX YCB is a benchmark dataset for hand object reconstruction. It contains over than 500K multi-view images from 10 objects grasping objects. The dataset provides both 3D hand and 2D keypoint annotations.
- ARCTIC [6] is a large scale dataset of bimanual hand-object manipulations containing over than 400K images from both egocentric and third-person views. It contains both single and dual hand manipulations along with accurate 3D and 2D hand annotations.
- Hot3D [2] is a recent egocentric dataset from daily activities that includes a high degree of occlusions and can significantly enhance the performance of WiLoR in egocentric scenarios. It contains both 3D hand and 2D keypoint annotations.

Additionally, we include four synthetic datasets that provide accurate ground truth 2D and 3D annotations:

- RHD [22] is amongst the first synthetic datasets of

hands rendered under different illumination patterns. The dataset is composed of 62K images with accurate 3D and 2D annotations.

- Re:InterHand [12] is synthetic dataset that extends InterHand2.6M by rendering hands under different illuminations and environments to bridge the gap between studio setup and in-the-wild images. The dataset provides both 3D hand and 2D keypoint annotations.
- BEDLAM [3] is a large scale full body synthetic dataset, that has proven extremely effective in whole body reconstruction tasks [4]. We use BEDLAM and randomly crop regions around the human hands to augment the training data. We use over 500K image crops. The dataset provides both 3D hand and 2D keypoint annotations.

Finally, we include in-the-wild datasets that contain only 2D information, but can effectively boost the generalization performance of WiLoR in in the while scenarios.

- COCO WholeBody [9] is a subset of COCO dataset and one of the main benchmarks for body pose estimation. It contains over than 100K in the wild images with humans participating in different activities. It provides 2D hand keypoint annotations.
- Halpe [7] is a full body in-the-wild dataset, composed of over than 40K images will 2D keypoint annotations. The dataset contains 21 keypoints for each hand.
- MPII NZSL [15] is a common benchmark for human body pose estimation, containing a set of in-the-wild, synthetic and lab environment images. The dataset contain 15K images with 2D keypoint hand annotations.

5. Temporal Coherence

In the project page we provide several videos that demonstrate the temporal coherence of WiLoR in challenging scenarios such as kneading dough or playing guitar. Despite being trained on single images, WiLoR can provide smooth reconstructions given its stable and robust detections.

References

[1] Vasileios Baltatzis, Rolandos Alexandros Potamias, Evangelos Ververas, Guanxiong Sun, Jiankang Deng, and Stefanos Zafeiriou. Neural sign actors: A diffusion model for 3d sign language production from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1985–1995, 2024. 1

[2] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 2

[3] Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8726–8737, 2023. 3

[4] Zhongang Cai, Wanqi Yin, Ailing Zeng, Chen Wei, Qingping Sun, Wang Yanjun, Hui En Pang, Haiyi Mei, Mingyuan Zhang, Lei Zhang, et al. Smpler-x: Scaling up expressive human pose and shape estimation. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2

[6] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J. Black, and Otmar Hilliges. ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2

[7] Hao-Shu Fang, Jiefeng Li, Hongyang Tang, Chao Xu, Haoyi Zhu, Yuliang Xiu, Yong-Lu Li, and Cewu Lu. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[8] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2

[9] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 196–214. Springer, 2020. 3

[10] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[11] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[12] Gyeongsik Moon, Shunsuke Saito, Weipeng Xu, Rohan Joshi, Julia Buffalini, Harley Bellan, Nicholas Rosen, Jesse Richardson, Mize Mallorie, Philippe Bree, Tomas Simon, Bo Peng, Shubham Garg, Kevyn McPhail, and Takaaki Shiratori. A dataset of relighted 3D interacting hands. In *NeurIPS Track on Datasets and Benchmarks*, 2023. 3

[13] Georgios Pavlakos, Dandan Shan, Ilija Radosavovic, Angjoo Kanazawa, David Fouhey, and Jitendra Malik. Reconstructing hands in 3D with transformers. In *CVPR*, 2024. 1

[14] Rolandos Alexandros Potamias, Stylianos Ploumpis, Stylianos Moschoglou, Vasileios Triantafyllou, and Stefanos Zafeiriou. Handy: Towards a high fidelity 3d hand shape and appearance model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4670–4680, 2023. 1

[15] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 3

[16] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 2

[17] Pengfei Xie, Wenqiang Xu, Tutian Tang, Zhenjun Yu, and Cewu Lu. Ms-mano: Enabling hand pose tracking with biomechanical constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2382–2392, 2024. 1

[18] Yufei Xu, Jing Zhang, Qiming Zhang, and Dacheng Tao. ViTPose: Simple vision transformer baselines for human pose estimation. In *Advances in Neural Information Processing Systems*, 2022. 1

[19] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9043–9053, 2023. 1

[20] Jinglei Zhang, Jiankang Deng, Chao Ma, and Rolandos Alexandros Potamias. Hawor: World-space hand motion reconstruction from egocentric videos. *arXiv preprint arXiv:2501.02973*, 2025. 1

[21] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 1

[22] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 2

[23] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2

[24] Ronglai Zuo, Rolandos Alexandros Potamias, Evangelos Ververas, Jiankang Deng, and Stefanos Zafeiriou. Signs as tokens: An autoregressive multilingual sign language generator. *arXiv preprint arXiv:2411.17799*, 2024. 1