Towards Accurate Alignment in Real-time 3D Hand-Mesh Reconstruction
Supplementary Material

Xiao Tang Tianyu Wang Chi-Wing Fu
The Chinese University of Hong Kong
{xtang,wangty,cwfu}@cse.cuhk.edu.hk

Overview

This supplementary document is composed of the following sections.

- In Section 1, more qualitative comparisons between our method and the state-of-the-arts are provided.
- In Section 2, more visual results produced by our method are provided.
- In Section 3, we report the alignment comparison result on the HO-3D dataset [1].
- In Section 4, we report the ablation study result on finger-level alignment.
- In Section 5, we show some very challenging cases that our method cannot handle.
- In Section 6, we provide description on the supplementary video, which features five AR scenarios.

1. Additional Results: Visual Comparisons on FreiHAND

Figure 1 shows more qualitative comparisons among Hasson *et al.* [2], I2L-MeshNet [3], and our method on the FreiHAND dataset [5]. From these results, we can see that our method predicts hand meshes which better match the hand gestures in the input images with higher quality finger-level alignment.

Figure 1. Visual comparisons between our method and state-of-the-arts. The input images are from the FreiHAND dataset [5].

2. Additional Results on EgoDexter and FreiHAND

Figure 2 shows more results produced by our method on the test images in the EgoDexter dataset [4] (unseen) and the FreiHAND dataset [5]. Our method is able to generate hand meshes that better match the hand gesture in the input images and better align with the real hand in the image space.

Figure 2. More qualitative results on the test images in the EgoDexter dataset (left) and in the FreiHAND dataset (right).

3. Quantitave Experiments: Alignment Comparison on HO-3D

In the main paper, we reported the alignment performance between our method and I2L-MeshNet [3] on the FreiHAND dataset. Here, we further selected 2,000 images from the training set of HO-3D dataset [1] and directly tested both methods on them without re-training. The quantitative result in Table 1 below shows that our method can predict hand meshes with better image-mesh alignment on unseen data, indicating a good generalizability of our method.

Table 1. Quantitative alignment comparisons between our method and I2L-MeshNet [3] on the HO-3D dataset [1]. ↓ means the lower the better and ↑ means the higher the better. We report the average HD for the fingers and palm in the evaluation of finger-level alignment.

Methods	Hand-level		Finger-level	
	mIoU ↑	HD ↓	mIoU ↑	HD ↓
I2L-MeshNet [3]	78.71	16.30	46.46	21.89
Our w/o ObMan [2]	79.02	15.93	46.84	21.53
Our w/ ObMan [2]	**80.64**	**14.67**	**50.95**	**20.15**

4. Ablation Study: Finger-level Alignment

In the main paper, we presented an ablation study of our network based on hand-level alignment. Here, we report the same ablation study but based on finger-level alignment in Table 2 below. Both ablation studies show that our full pipeline (with all components) performs the best.

Table 2. Comparing the performance of our full pipeline (bottom-most) with various ablation cases. Note that we report the average HD for the fingers and palm in the evaluation of finger-level alignment. The experiment is conducted with three runs.

Methods	Hand-level		Finger-level	
	mIoU ↑	HD ↓	mIoU ↑	HD ↓
w/o joint stage	92.42±0.10	5.12±0.08	76.11±0.20	7.21±0.19
w/o refine stage	92.05±0.04	5.93±0.07	75.83±0.18	7.38±0.20
w/o local	92.55±0.05	5.34±0.04	76.22±0.14	7.15±0.18
w/o global	92.63±0.04	5.01±0.07	76.37±0.14	7.09±0.15
w/o offset	88.01±0.10	7.97±0.11	67.06±0.21	10.46±0.24
w/o $\mathcal{L}_{\text{sil}}$	92.50±0.08	5.19±0.08	75.88±0.15	7.28±0.15
w/o $\mathcal{L}_{\text{render}}$	92.80±0.02	4.97±0.04	76.36±0.10	7.00±0.11
Full	**92.95**±0.04	**4.70**±0.02	**77.22**±0.16	**6.82**±0.17

5. Challenging Cases

Figure 3 below shows three very challenging cases that our method cannot handle: (a) when a large portion of the hand is occluded (say by the wrist and arm) in the image; (b) when the image exhibits severe motion blur; and (c) when the hand gesture is rare.

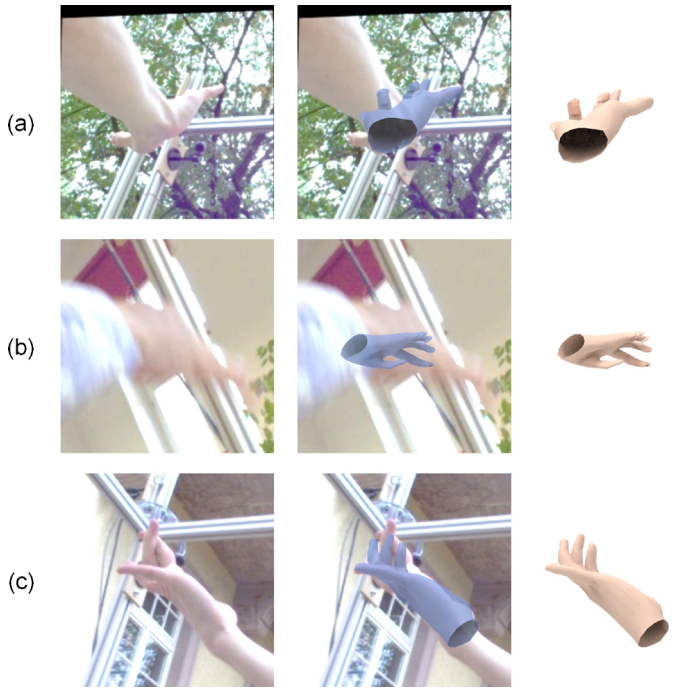

Figure 3. Three very challenging cases that our method cannot handle.

6. Supplementary Video

Lastly, we showcase five AR scenarios in our supplementary video, demonstrating the applicability of our method: (i) grab a virtual 3D ping-pong paddle and move it; (ii) wrap a virtual paper band around user's hand; (iii) try on a virtual ring; (iv) interact with a virtual water simulation in 3D; and (v) grab and manipulate a virtual walkie-talkie, e.g., pressing a button on it.

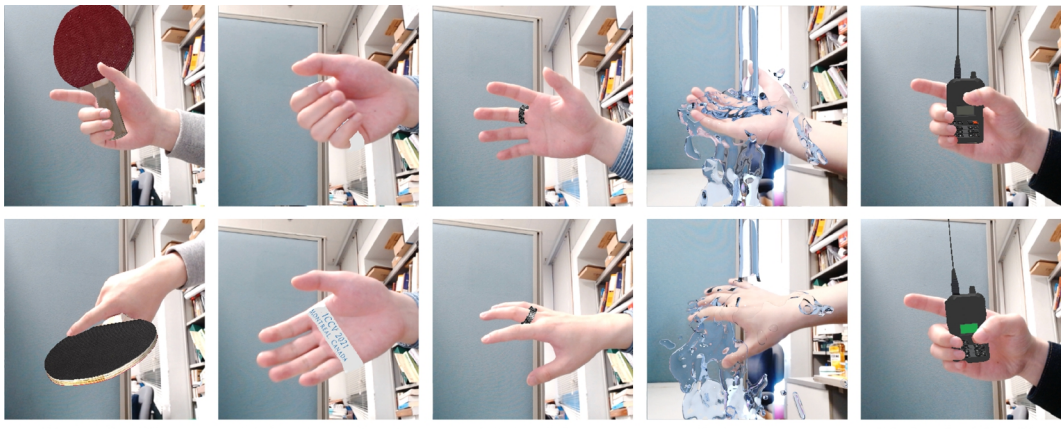

Figure 4. Five example AR scenarios developed based on our method. Please refer to the supplementary video.

References

[1] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 1, 4

[2] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *ICCV*, pages 11807–11816, 2019. 2, 4

[3] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, pages 752–768, 2020. 2, 4

[4] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In *ICCV*, pages 1154–1163, 2017. 3

[5] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for marker-less capture of hand pose and shape from single RGB images. In *ICCV*, pages 813–822, 2019. 2, 3