Decoupled Iterative Refinement Framework for Interacting Hands Reconstruction from a Single RGB Image (Supplemental Material)

In the supplemental material, we provide:

- more details of network structure and computational requirements in Sec. 1,
- the details of mesh smooth loss in Sec. 2,
- more quantitative results in Sec. 3,
- more ablation experiments in Sec. 4,
- qualitative results in real scenarios in Sec. 5,

Note that all the notation and abbreviations here are consistent with the main manuscript.

1. Details of Network Structure and Computational Requirements

In this section, we introduce the structure details of the feature fusion layers, the Graph Convolutional Network (GCN) and the transformers. First, we use a residual convolutional module as the stacked hourglass network [6] to fuse the feature maps from the encoder and previous decoding layer. Meanwhile, we use a residual convolution module to enhance the fused visual feature map with projected features. We set the number of channels of fused visual features and enhanced visual features to 256. We adopt a 4-layer semantic GCN [11] to perform information interaction between the single-hand joints, where the number of channels of joint features is 128. We adopt a 4-layer transformer [12] for information interaction between two-hand joints, in which we add spatial position encoding to the input joint features. Wtih a single GPU (NVIDIA A100) and a batch size of 64, for the network with two refinement stages, the training time is 39.8h, the memory usage is 22.1G, the FLOPs is 30.8G, and the model parameters are 55.1M.

2. Mesh Smooth Loss

Following previous methods [9, 4, 3], we use mesh smooth loss to maintain the estimated mesh geometry reasonable, including a normal consistency loss L_{norm} and edge length consistency loss L_{edge}. L_{norm} is defined as follows:

$$L_{norm} = \sum_{f} \sum_{\{i,j\} \subset f} \left\| \left\langle \mathbf{e}_{ij}, \mathbf{n}_{f}^{gt} \right\rangle \right\|_{1}, \tag{1}$$

where f and $\mathbf{n}_f$ indicate a face of the hand mesh and the unit normal vector of face f, respectively. $\mathbf{e}_{ij}$ indicates a edge of the face f. $\langle \cdot, \cdot \rangle$ is the inner product of two vectors.

L_{edge} is defined as follows:

$$L_{edge} = \sum_{f} \sum_{\{i,j\} \subset f} \left\| \left\| \mathbf{e}_{ij} \right\|_{2}, \left\| \mathbf{e}_{ij}^{gt} \right\|_{2} \right\|_{1}. \tag{2}$$

L_{edge} constrains each edge of the predicted mesh to have the same edge length as the ground truth.

3. More Quantitative Results

In the dataset proposed by Tzionas et al. [8], we selected the sequence containing two hands for quantitative experiments. In particular, we only use this dataset for testing and all models are trained using the InterHand2.6M [5]. As shown in Table 1, our method outperforms IntagHand [3] and InterShape [10] by a large margin on [8], which further demonstrates the superior generalization ability of our method.

4. Ablation Study

Our basic model regresses the hand model parameters directly from the visual features. We tried multiple good practices to improve the performance of the basic model,

	MPJPE	MPVPE
InterShape [10]	18.24	17.93
IntagHand [3]	19.21	18.91
Ours	**16.48**	**16.25**

Table 1. Quantitative results on Tzionas et al. [8]. We report the MPJPE (mm) and MPVPE (mm).

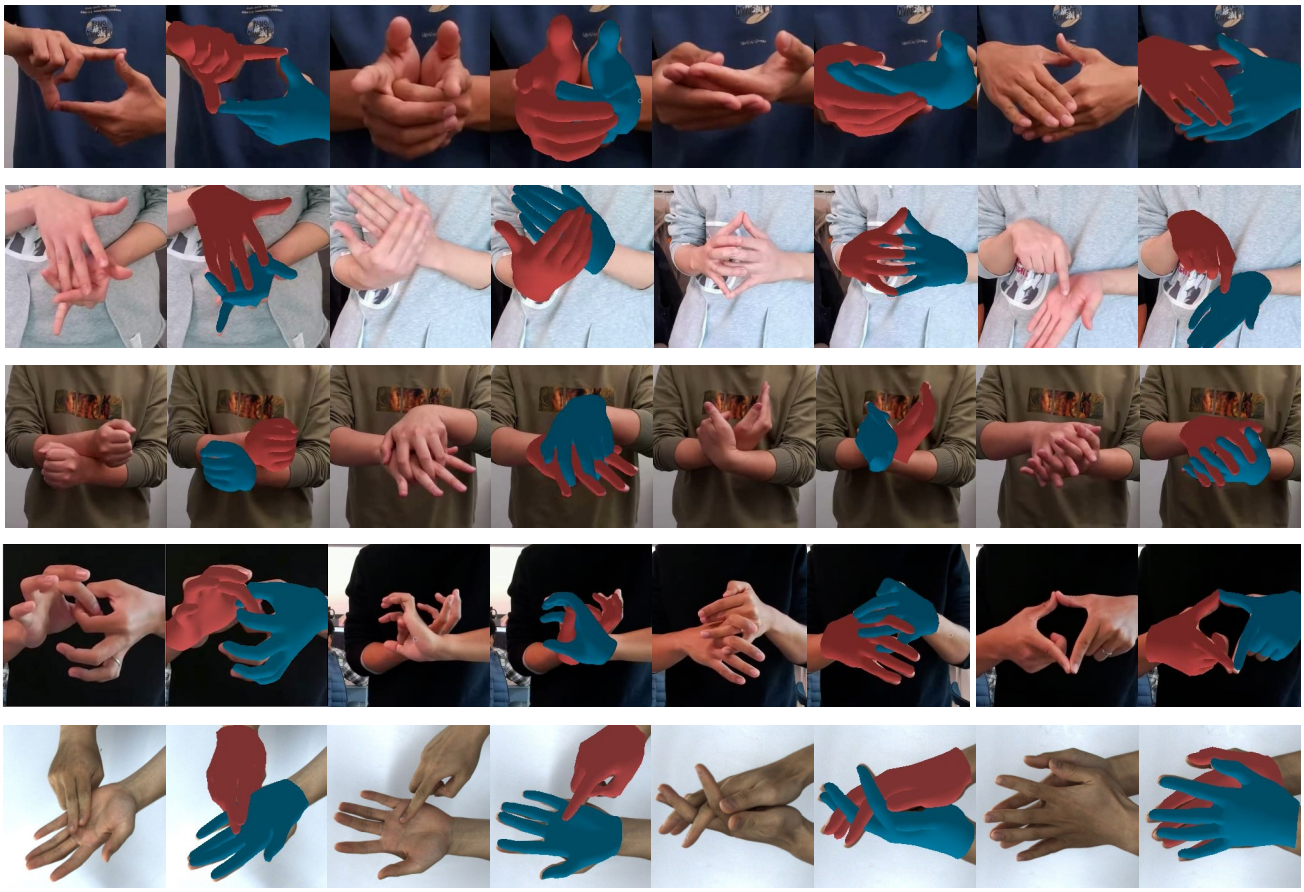

Figure 1. Two-hand reconstruction results in real scenarios on different subjects with different hand shapes and hand poses.

Method	MPJPE	MPJPE	MIAA
Baseline	12.44	12.11	7.41
w/o Attn	12.53	12.23	7.53
w/o Smooth L1	12.49	12.20	7.50
w/o Motion Blur	12.50	12.21	7.51
w/o large LR	12.79	12.45	7.64
w/o All	13.01	12.67	7.83

Table 2. Ablation study of the basic model on InterHand2.6M [5]. We report the MPJPE (mm), MPVPE (mm) and MIAA (pixel).

including adopting a larger learning rate (from 1e-4 to 3e-4), adopting SmoothL1 loss [1, 7] instead of L1 loss, adopting motion blur for data augmentation, adopting the attention mechanism to obtain the different feature of the left and right hand respectively, etc. As shown in Table 2, using the attention mechanism, data augmentation and Smooth L1 loss have an impact of about 0.1 mm on the basic model, and the use of a larger learning rate has an impact of close to 0.35 mm. If these components are removed, the MPJPE drops to 13.01 mm. It is worth mentioning that compared to the previous methods, our method can adopt a larger learning rate due to the simple and efficient network design.

5. Qualitative Results

As shown in Fig. 1, we experiment on five subjects in real scenarios. The five subjects have different hand shapes and hand poses. First, our method is able to generate relatively accurate mesh-image alignments for unseen subjects. Second, our method can also perform reasonable reconstructions for some unseen complex interacting poses. Overall, our method achieves efficient pixel-level alignment and 3D spatial relationship modeling thanks to the decoupled design of 2D visual feature space and 3D pose feature space. At the same time, sparse and compact node-level information interaction avoids overfitting and achieves strong generalization ability. In particular, we provide a video in the Supplementary Materials to demonstrate the strengths of our method for spatial relationship modeling and image-mesh alignment compared to SOTA method IntagHand [3].

References

[1] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 2

[2] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, pages 11090–11100, 2022.

[3] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, pages 2761–2770, 2022. 1, 2

[4] 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. Springer, 2020. 1

[5] 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 *ECCV*, pages 548–564. Springer, 2020. 1, 2

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, pages 483–499. Springer, 2016. 1

[7] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 2

[8] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 1

[9] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018. 1

[10] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *CVPR*, pages 11354–11363, 2021. 1

[11] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019. 1

[12] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estimation with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021. 1