# Reconstructing Interacting Hands with Interaction Prior from Monocular Images
# —Supplementary Material—

Binghui Zuo[1] Zimeng Zhao[1] Wenqian Sun[1] Wei Xie[1] Zhou Xue[2] Yangang Wang[1]*

[1]Southeast University, China [2]Pico IDL, ByteDance, Beijing

In this supplementary document, we first give more descriptions of the proposed dataset *Two-hand 500K* in Sec. A. Then, we separately introduce how to overlay mesh on the image in Sec. B and more details of training loss in Sec. C. Finally, we show more comprehensive evaluations of our method in Sec. D.

## A. Two-hand 500K

### A.1. Background.

Two-hand datasets are fewer than single-hand datasets. Most related works rely on the well-known Interhand2.6M [6] in an images-paired training manner. Even with a large scale, there are less than 10K interaction states in [6], which is unfriendly for building an expressive interaction prior. Different from prior works, we skillfully break the dependency on images-paired data and train prior with multimodal datasets. We also propose *Two-hand 500K* , a large-scale dataset focusing on two-hand interaction. *Two-hand 500K* plays a positive effect on expanding the interaction states. We collect *Two-hand 500K* in two ways: MoCap and Splicing.

### A.2. Data from MoCap.

To accurately and quickly collect interaction states, we use a marker-based MoCap system to capture skeleton data, as well as the automatically annotated joint positions. Fig. 1(a) shows MoCap equipment that contains wearable VR eye, movable handles and interactive gloves. Among them, the handles are used for global positioning, while the gloves are used for local positioning. There are six wireless inertial sensors integrated into each glove, which makes finger poses more precise. We show our capturing process in Fig. 1(b). When capturing, to ensure diversity, we imitate interaction patterns from related media on the Internet and display gestures for daily communication such as greeting, sign language and laboring *et al*. The captured skeleton data is demonstrated in Fig. 1(c), where each hand contains 16 joints. We also fit MANO parameters [7] using the inverse kinematics algorithm (IK) to give *Two-hand 500K* a broader application as shown in Fig. 1(d). The physics engine is adopted to optimize the fitted MANO parameters to make the interaction more physically plausible, which is shown in Fig. 1(e).

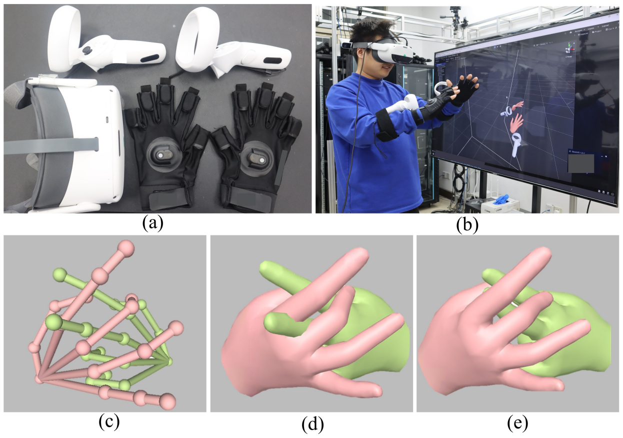


Figure 1. **MoCap process**. (a) Overview of our marker-based MoCap system; (b) Capturing skeleton data; (c) The captured skeleton data has 16 joints for each hand, all of which are annotated automatically; (d) The fitted MANO parameters; (e) Physically plausible interaction.

### A.3. Data from Splicing.

To erase the barrier between the single-hand and two-hand datasets, we splice single hands to generate interaction. Specifically, we pick out five single-hand datasets [6, 12, 2, 13, 10] that contain both left and right hands. We randomly sample left-hand frames and right-hand frames from these datasets and randomly splice them with random rela-

*Corresponding author. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), in part by the Natural Science Foundation of Jiangsu Province (No. BK20220127).

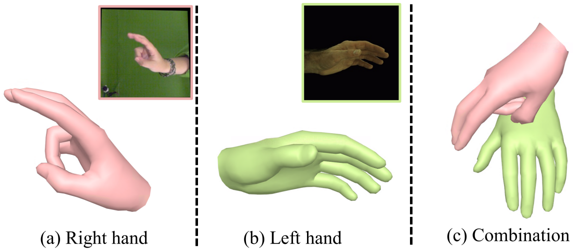


Figure 2. **Splicing process**. (a) Right hand is sampled from [13]; (b) Left hand is sampled from [6]; (c) Combined interaction.

tive translations. To ensure that the combinations are physically plausible and encourage proximity, we also adopt the physics engine to refine interaction. Fig. 2 shows the splicing process from two separate hands to a combination. The right hand is sampled from [13], while the left hand is sampled from [6].

**Implementation details.** We use *Bullet* [1] as our physics simulation platform, which is used throughout the paper. Similar to [11], we convert each hand mesh to 16 articulated ellipsoids for better collision detection. Considering that entanglement between two hands can lead to a long time without reaching the target interaction states, we filter out the states if the target interaction is not completed within 2.0s. More details can be found in [11]. In addition, when collecting with MoCap, we set the sampling frame rate to 25fps, which is up to 120.

### A.4. Quantitative evaluations for Two-hand 500K.

We also add additional quantitative evaluations for the proposed *Two-hand 500K* and Interhand2.6M [6] in Tab. 1, including the interaction volume (*Inter.V*), penetration depth (*Pene.D*) and contact ratio (*Cont.R*). Benefiting from the optimization of the physics engine, the collected interaction states are more physically plausible.

| Dataset | Inter.V($cm^3$)↓ | Pene.D(cm) ↓ | Cont.R(%)↓ |
|---|---|---|---|
| Interhand2.6M | 4.07 | 0.26 | 43.61 |
| Two-hand 500K | 1.04 | 0.11 | 16.51 |

Table 1. **Quantitative evaluations for Interhand2.6M [6] and *Two-hand 500K*** . All the evaluations show that the proposed dataset provides higher-quality interaction data.

## B. Mesh Overlay

Since the joint coordinates of the interacting hands predicted by our VAE decoder are based on the root of the right hand as the reference system, we align this estimated 3D result with the input image according to the following steps. Considering that two-hand interaction is usually captured at a close position to the camera, orthogonal projection is used to model the camera in this process:

$$\begin{bmatrix} u_{\mathrm{j}} \\ v_{\mathrm{j}} \\ 1 \end{bmatrix} = \frac{1}{z_{\mathrm{j}}} \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot [\mathbf{R}|\boldsymbol{t}] \cdot \begin{bmatrix} x_{\mathrm{j}} \\ y_{\mathrm{j}} \\ z_{\mathrm{j}} \end{bmatrix}, \qquad (1)$$

where $s$ is a scaling factor to convert the hand scale (*e.g.* meter unit for MANO) to pixel unit. $[\mathbf{R}|\boldsymbol{t}]$ is the camera extrinsic parameter, and $\mathbf{R}$ is fixed as identity matrix. $(u_{\mathrm{j}}, v_{\mathrm{j}})$ is estimated as the local maximum of the corresponding $\mathbf{H}_{\mathrm{j}}$. Considering that $\{\mathbf{H}_{\mathrm{j}}\}_{\mathrm{j}=1}^{42}$ do not attempt to obtain high positioning accuracy, those maps with a single peak are given greater weights.

## C. Loss Functions

Different hand pose representations can be estimated by our framework. In this section, we introduce more details of the loss functions that are used for these representations. In the following, symbols with the hat superscripts refer to prediction, while the star superscripts refer to ground truth.

### C.1. Representations of 3D joints.

To encourage the accuracy of 3D hand joints, we use L1 distance to supervise the 3D joint positions between prediction and ground truth.

$$L_{\mathrm{j}} = \sum_{h} \sum_{i=1}^{21} \left\| \hat{\boldsymbol{J}}_{h,i} - \boldsymbol{J}^{*}_{h,i} \right\|_{1}, \qquad (2)$$

where $h \in \{\text{right, left}\}$ denotes right and left hand, $i$ refers to the index of hand joints.

### C.2. Representations of 3D vertices.

The same loss term as 3D joints is adopted to supervise 3D hand vertices, which is defined as follows:

$$L_{\mathrm{v}} = \sum_{h} \sum_{i=1}^{778} \left\| \hat{\boldsymbol{V}}_{h,i} - \boldsymbol{V}^{*}_{h,i} \right\|_{1}, \qquad (3)$$

where $h \in \{\text{right, left}\}$ denotes right and left hand and $i$ refers to the index of hand vertices.

### C.3. Representations of MANO parameters.

We use MSE loss between the estimated MANO parameters and ground truth. For this representation, both hand pose and shape parameters are embedded into the interaction prior.

$$L_{\mathrm{MANO}} = \left\| \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^{*} \right\|_{2}^{2} + \left\| \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{*} \right\|_{2}^{2}. \qquad (4)$$

**Normal loss.** Besides the above, we also use the normal loss as [4, 5] to promote the smoothness and reasonableness of reconstruction. For a predicted triangular mesh,

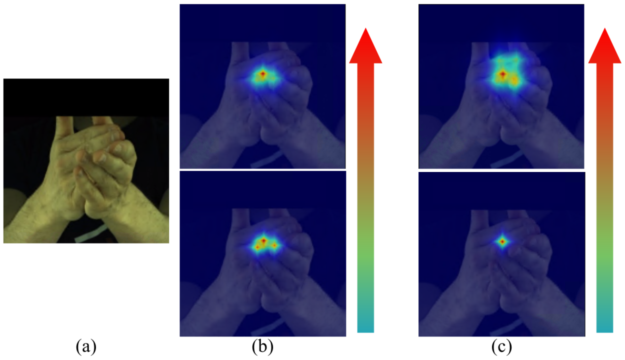


Figure 3. **Effectiveness of different hyperparameters**. (a) Input; (b) IAH with different variance $\sigma$, from bottom to top, variance becomes larger; (c) IAH with different adjacency $d$, from bottom to top, adjacency becomes larger.

the normal perpendicular to three adjacent edges should be equally perpendicular to the corresponding three edges in the ground truth.

$$L_{\text{nor.}} = \sum_{f=1}^{F} \sum_{\{i,j\}\in f} \left\| \left\langle \frac{\boldsymbol{V}_i^f - \boldsymbol{V}_j^f}{\left\| \boldsymbol{V}_i^f - \boldsymbol{V}_j^f \right\|_2}, n_f^* \right\rangle \right\|_1, \qquad (5)$$

where the number of $F$ is 1552, a watertight face based on the MANO template. $i, j$ denote the index of the vertices $\boldsymbol{V}$, both $i$ and $j$ belong to the same face of the triangular mesh. We compute the unit normal vector $n_f^*$ from the ground truth.

**Penetration loss.** For the close interaction like entangled fingers, even with the constraints mentioned above, penetrations still exist. We adopt a penetration loss term based on differentiable Signed Distance Field (SDF) [3, 8] to punish penetration. According to the definition of the SDF, we know that a positive value is assigned if a vertex $(x, y, z)$ is inside the hand surface. On the contrary, we set the SDF to zero.

$$L_{\text{pene.}} = \sum_{h} \sum_{v=1}^{778} - \min(\text{SDF}(x, y, z), 0), \qquad (6)$$

where $h \in \{\text{right, left}\}$, meaning that we compute SDF values for the left and right hand independently.

# D. More Experiments

## D.1. More qualitative results.

Besides the results reported in the main submission, we demonstrate more results in this section to prove the satisfactory performance of our method. As shown in Fig. 7, we report more qualitative results on Interhand2.6M [6]. Even for those interaction states with heavy self-occlusion or complex entanglement, the reconstruction quality is still pretty good. Apart from that, we also report performance on [9]. Although we do not train the network on it, the reconstructed interactions are also realistic. We show the results in Fig. 4.

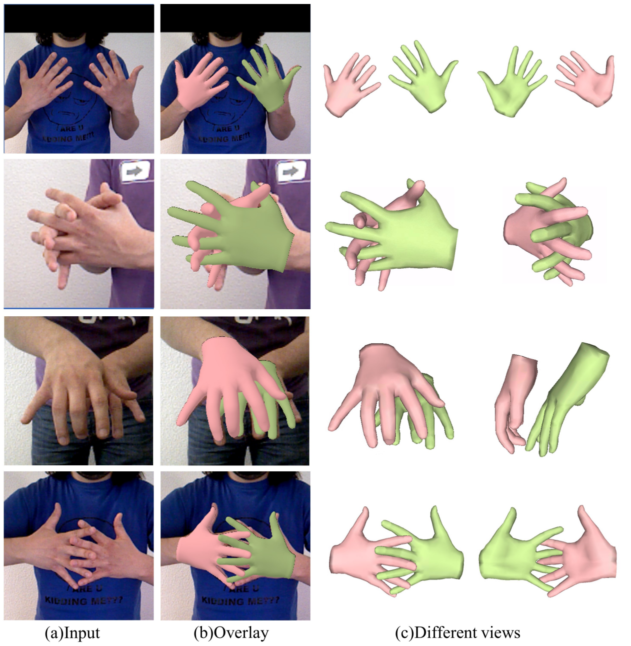


Figure 4. **More results on [9]**. (a) Input images; (b) Mesh overlay; (c) Different views of reconstructed interaction.

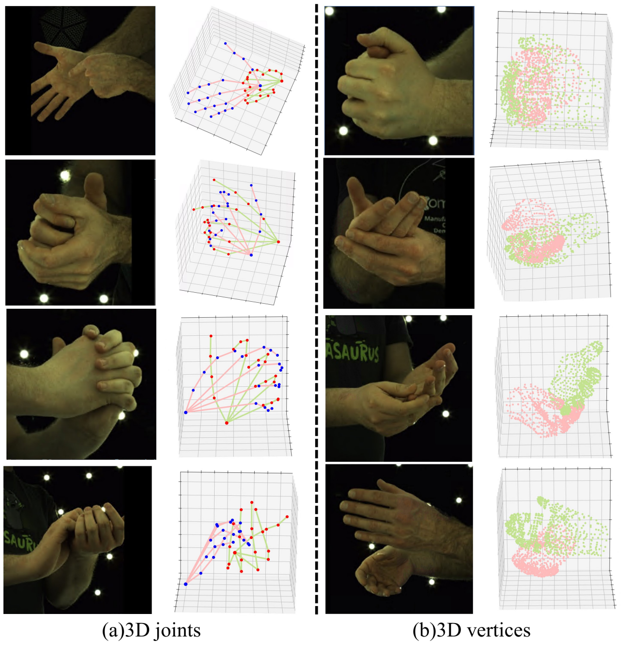


Figure 5. **Different hand pose representations**. (a) Hand pose representations of 3D joints; (b) Hand pose representations of 3D vertices.

| $\sigma$ | $\alpha$ | $d$ | MPJPE $\downarrow$ | MPVPE $\downarrow$ |
|---|---|---|---|---|
| 2.0 | 2.0 | 3.0 | 8.91 | 9.12 |
| 3.0 | 2.0 | 2.5 | 8.67 | 8.84 |
| 2.0 | 2.5 | 2.5 | 8.53 | 8.69 |
| 2.0 | 2.0 | 2.5 | 8.34 | 8.51 |

Table 2. **Ablation study of the hyperparameters**. We report the influence of hyperparameters variance $\sigma$, zoom factor $\alpha$ and adjacent regions $d$.

### D.2. More ablation studies.

In the main submission, we have introduced the generation and effectiveness of IAH. With different variances and adjacent regions, the hand joints can be adaptively mapped to a certain heatmap, where the resolution is fixed in 64×64. As shown in Fig. 3 (b), with a larger zoom factor $\alpha$, the distribution located in adjacent joints is fuzzier than in the identity joint. Similarly, the size of adjacent region $d$ also affects the expression of IAH. Because a larger $d$ causes joints in non-interactive areas to be mapped, as demonstrated in Fig. 3 (c). To maximize the performance of IAH, we ablate the effect with different hyperparameters and set two candidates for each hyperparameter, as shown in Tab. 2. The most suitable set of hyperparameters is listed in the last row, which balances the joints in adjacency and distribution on them.

### D.3. Different pose representations.

The proposed framework is compatible with the different hand pose representations. As mentioned in the main submission, three hand pose representations are considered in our framework. Besides the representation of MANO parameters summarized above, we show the performance when estimating hand joints and vertices. On the left side of Fig. 5, we report the evaluations on hand joints. At the same time, the right side of Fig. 5 reflects the performances on hand vertices. All of these representations have achieved convincing results.

### D.4. Failure cases.

We list two failure cases in Fig. 6, where our method cannot reconstruct reasonable interaction. We attribute the failure to the inappropriateness of the extracted features, which leads to incorrect sampling from the constructed interaction prior.

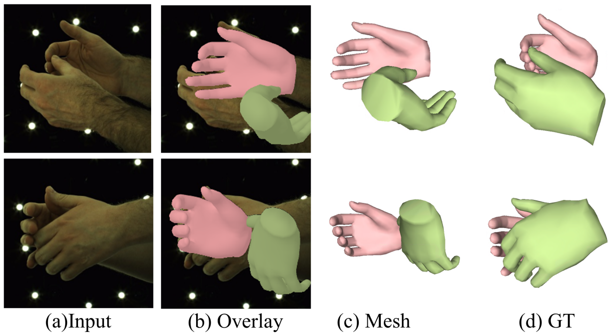


Figure 6. **Failure cases on [6]**. (a) Input; (b) Overlay mesh on the images; (c) Reconstructed interaction; (d) Ground truth.

## References

[1] Bullet. https://github.com/bulletphysics/bullet3. 2

[2] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019. 1

[3] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 3

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2761–2770, 2022. 2

[5] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 2

[6] 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*, pages 548–564. Springer, 2020. 1, 2, 3, 4, 5

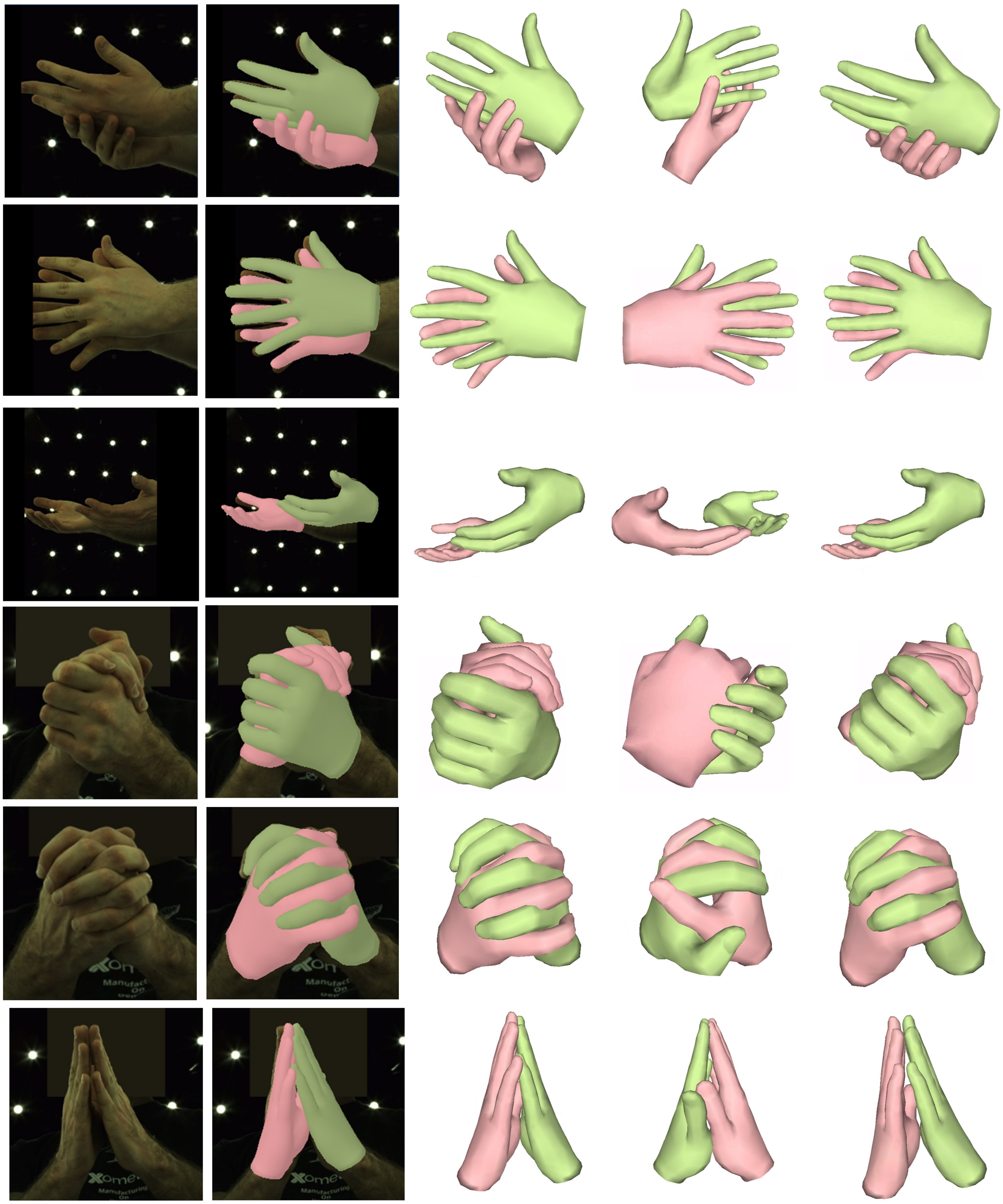
[7] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Gr.*, 36(6):1–17, 2017. 1

[8] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *2021 International Conference on 3D Vision (3DV)*, pages 432–441. IEEE, 2021. 3

[9] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 3

[10] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 1

[11] Zimeng Zhao, Binghui Zuo, Wei Xie, and Yangang Wang. Stability-driven contact reconstruction from monocular color images. In *Proceedings of the IEEE/CVF Conference on*



(a)Input (b)Overlay (c)Different Views (d)GT

Figure 7. **More results on [6]**. Our method received satisfactory reconstruction in most scenarios. Even for strong self-occlusions, the pose is still accurate.

*Computer Vision and Pattern Recognition*, pages 1643–1653, 2022. 2

[12] 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. 1

[13] 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. 1, 2