RenderIH: A large-scale synthetic dataset for 3D interacting hand pose estimation (*Supplementary Material*)

Lijun Li*[1,2], Linrui Tian[1], Xindi Zhang[1], Qi Wang[1], Bang Zhang[1], Liefeng Bo[1], Mengyuan Liu[3], and Chen Chen[4]

[1]Alibaba Group, [2]Shanghai Artificial Intelligence Laboratory, [3]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, [4]Center for Research in Computer Vision, University of Central Florida

This supplementary material contains additional information that could not be included in the main manuscript due to space limitations. We will begin by providing more detailed information about the dataset. Following that, we will briefly discuss the pose optimization details in our approach. Then we will then present additional visualization results from our qualitative experiments. Finally, we will discuss the broader impacts and limitations of our dataset.

A. More details on RenderIH

RenderIH is composed of 1 million synthetic images by varying the pose, camera view, and environment (texture, lighting, and background). By collecting annotations from IH2.6M, we removed samples of similar poses resulting in 3680 distinctive poses. For each distinctive pose, we augment $I = 30$ poses. After augmenting and optimization, we filter out those IH poses that still have notable penetration or exceed joint limits, the remaining data accounts for 93% of the total, and we produce approximately 100K natural and non-interpenetration IH poses. Then we apply 10 camera viewpoints to each pose and produce 1M synthetic images in total. For each image, we randomly pick from a collection of 300 HDR images to illuminate the hand and provide the background together with a hand texture map. The rendering process took more than 200 hours using 4 NVIDIA A100 GPUs. As for the corresponding annotation, we provide pose and shape parameters, 3D joint coordinates, 2D joint coordinates, and camera intrinsic and extrinsic parameters. It is worth noting that the synthetic data labels can be freely extended based on the user's preferences, such as generating hand parts segmentation masks. The automatically generated annotations are free of noise and are more flexible than the traditional labels of the real dataset. Some rendering examples to illustrate our photo-realistic effect are provided in the **video demo**.

*Corresponding author: 4065156@qq.com

a) Our penetration loss **b) Original multi-person loss**

Figure 1. Visual comparison of penetration loss between ours and multi-person penetration loss [1]. The penetration position of each hand can be obtained by utilizing the penetration loss. The green color and red color are used to denote where the right hand and left hand are penetrated respectively.

B. Pose optimization details

Penetration loss comparison. Hands are highly articulated and have a curved and concave shape as a whole object. It makes the original multi-person penetration loss [1] hard to detect the correct penetrated positions. However, by dividing the hand into 16 almost convex parts, we can perform SDF on each part and calculate the penetration, which can lead to more accurate penetration detection results. The visual difference between our algorithm and multi-person penetration loss [1] can be seen from Figure 1.

Optimization details. As shown in Figure 2 a), The anchor pairs are built between the closest anchors on both hands, making the IH has more contact. As shown in Figure 2 b), to avoid abnormal anchors pairs, the pair can only be established when $\bar{n_i^a} \cdot \bar{n_j^a} < 0$, in which $\bar{n^a}$ is the mesh face normal vector of the anchor. However, the IH attraction might cause a negative influence when the parts are in serious overlaps, as shown in Figure 2 c), there are conflicts

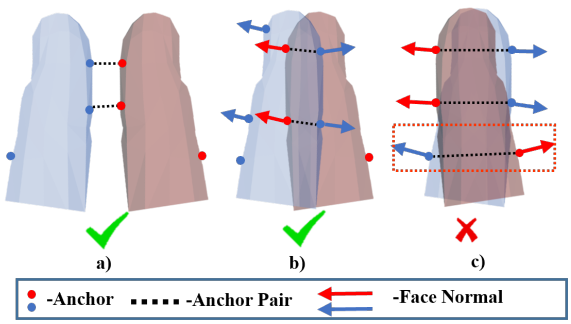

Figure 2. Building anchor pairs between IH. a) Nearest anchors are connected as the pairs, they tend to make more contacts between both hands. b) Using face normals to avoid abnormal pairs. c) The proper pairs are hard to build when the hand parts are in serious overlaps.

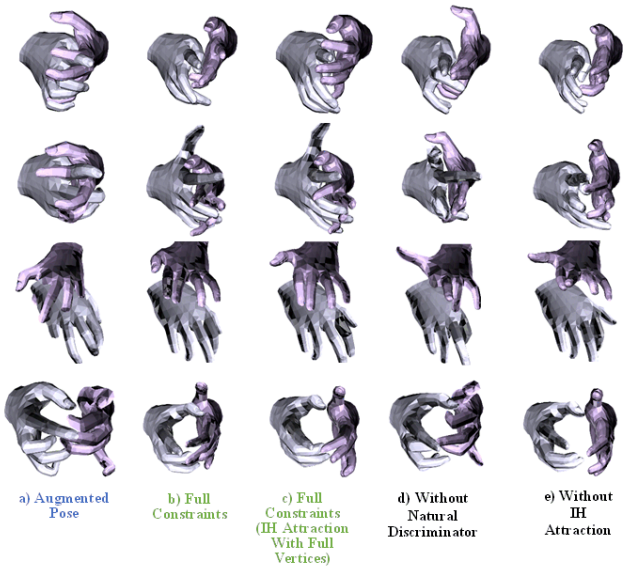

Figure 3. Results for different components of optimization. a) The poses augmented from the IH2.6M. b) After being optimized by the proposed method, the poses become valid and natural. c) Using vertices instead of anchors to make IH attraction has no significant differences. d) Poses optimized without discriminator are valid but not natural. e) Poses optimized without IH attraction have fewer contacts.

between pairs, making the mesh hard to separate, the simple way to solve this problem is separating the hands at first so that we could have better anchor pairs.

$$\underset{\psi^r,\psi^l}{argmin}(w_1 \sum_{i=1}^{A_r}\sum_{j=1}^{A_l} L_{ij}^A + w_2 L_a + w_3 L_{adv} + w_4 L_p), \quad (1)$$

In our implementation, we optimize the loss function in Equation 1 which is defined in the main paper in 215 iterations, we assign a larger weight w_4 for L_p and a smaller weight w_1 for L^A at the beginning to separate the hands, w_1 will increase while w_4 decrease during the optimization until 165th iteration. The anchor pairs will be rebuilt every 40 iterations to adapt to dynamically changing IH. The learning rate is set to 0.01 and will reduce after 20 no-loss-decaying iterations. Adam solver is utilized for optimization.

C. More visualization results

C.1. Results for different optimization components

Visualization of the effect of different components. We define multiple optimization loss functions to get valid and natural IH poses. As shown in Figure 3, the "Augmented Pose" is randomly augmented from the raw poses in IH2.6M, the joint poses are restricted according to Table 2 in the main paper. After being optimized by the full constraints, we get natural and non-interpenetration poses. Comparing Figure 3(b) and Figure 3(c), we can see that adopting anchors to make IH attraction has no significant differences from employing full vertices while reducing the time complexity. Furthermore, as demonstrated in Figure 3(d), the natural discriminator $\mathcal{D}$ could make the IH more natural, the **natural** poses are defined in the main paper, they not only conform to the anatomy but also frequently occur in daily life. Additionally, as shown in Figure 3(e), IH attraction enhances hand contact, which is hard to annotate in reality due to inter-occlusion.

Metrics	PAMPJPE/MPJPE/SMPJPE/MRRPE↓
Training set \ Test set	Tzionas
IH2.6M	6.76/16.78/13.97/14.63
IH2.6M+RenderIH	5.79/15.78/12.16/14.15

Table 1. The comparison of training with or without our dataset on IH2.6M dataset. Wrist joint is used as root.

C.2. Qualitative results comparison

Comparison with IntagHand. To better demonstrate the superiority of our data and method, we compare our result with the existing state-of-the-art method IntagHand [2] (Their models is also trained on the combination of IH2.6M [3] and synthetic images). Some qualitative comparisons with IntagHand are shown in Figure 6. By directly projecting 3D hand mesh onto the image, we can see our result is closer to the pose in the raw image. Additionally, the results of these images from various views are also presented (see Figure 4). In the first row of Figure 4, our result can be even better than the ground truth, where the middle, ring, and little fingers of the right hand are curved. To further compare our generalization ability, we compare with IntagHand on in-the-wild images (see Figure 5). The results show that our method can clearly achieve less inter-

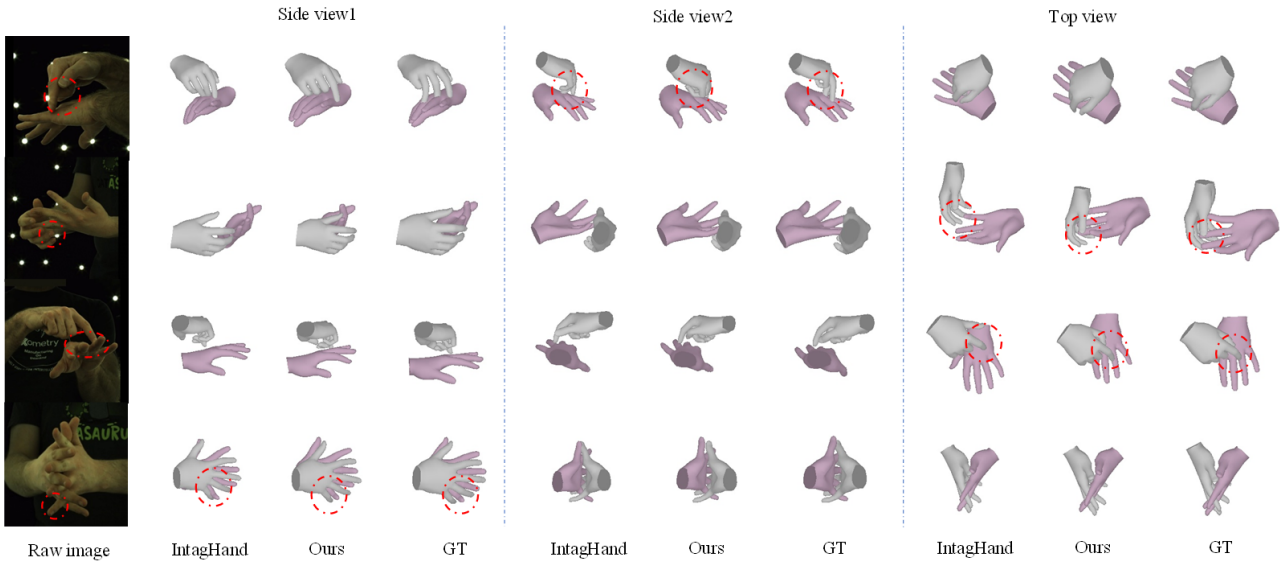

Figure 4. Qualitative comparison with our method and IntagHand [2] on InterHand2.6M under a variety of viewpoints and different levels of inter-hand occlusion. Red circles are used to highlight the positions where our methods can generate better results. In the first row, our result can be even better than the ground truth, where the middle, ring, and little fingers of the right hand are curved.

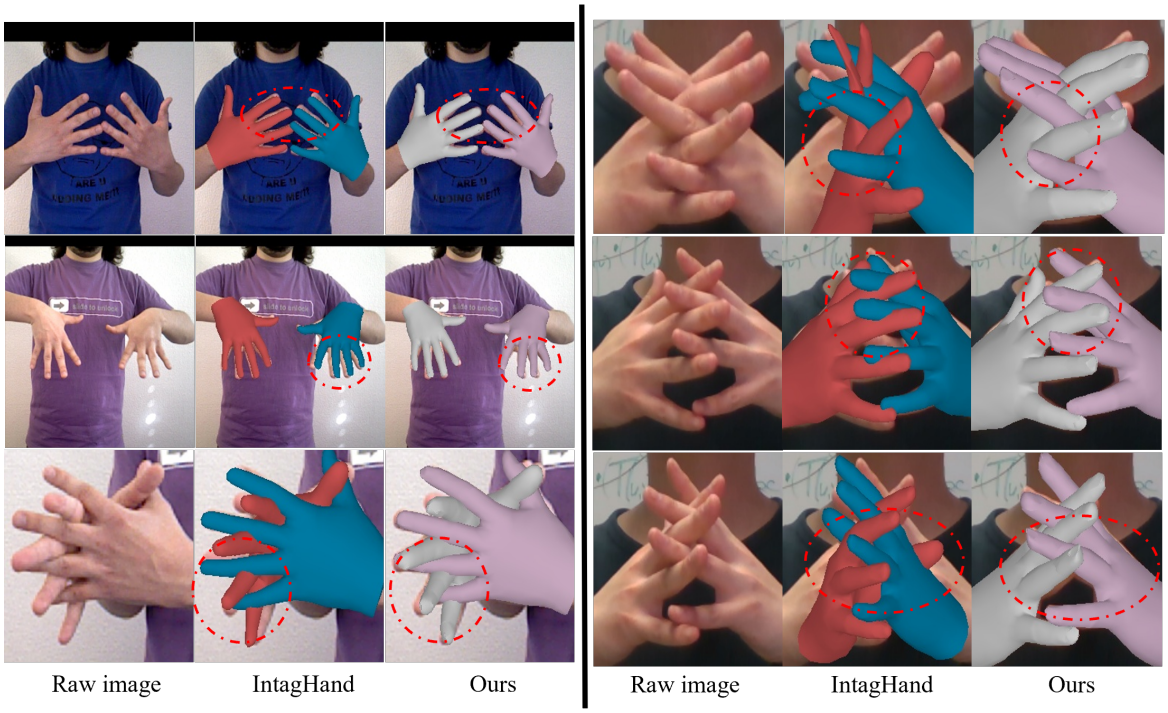

Figure 5. Qualitative comparison of the interacting hand pose estimation with our method and IntagHand [2] on in-the-wild images. The first three columns display images from the Tzionas dataset [4], while the last three columns showcase images from the RGB2Hands dataset [5]. Red circles are used to highlight the positions where our methods can generate better results. From the visualization results, we can clearly see that our model can generalize better for in-the-wild images.

penetration of two hands and more accurate finger interactions.

Impact of synthetic data. When only RenderIH is used for training, the performance is worse than when only IH2.6M is used, in part because the background variation in Tzionas is limited. The trend can be seen in the qualita-

Raw image IntagHand Ours

Figure 6. Qualitative comparison of the interacting hand pose estimation with our method and IntagHand [2] on InterHand2.6M. Red circles are used to highlight the positions where our methods can generate better results.

Raw image IH2.6M Only RenderIH Only RenderIH+IH2.6M

Figure 7. Qualitative comparison of our models trained on different training sets which test on in-the-wild images.

tive result in Figure 7. However, as a synthetic dataset, the function of our dataset is to largely reduce the number of real data needed for training instead of replacing real data entirely.

C.3. Quantitive results with wrist joint as root joint

For convenient future comparison, we report our model's performance using wrist joint as root joint following common practice. As shown in Table 1, the model trained on a mixture of RenderIH and IH2.6M demonstrates consistent improvement across all metrics compared to training on IH2.6M alone.

D. Broader impacts and limitations

Broader impacts. In this paper, we introduce a synthetic 3D hand dataset, RenderIH, with accurate and diverse poses. Since there are no large-scale synthetic interacting hand datasets, RenderIH will be impactful for the community, due to its unprecedented scale, diversity, and rendering quality. Moreover, the dataset not only can be used to improve the generalization ability in real scenes but also can be used for domain adaptation.

Limitations. The hyperparameters of pose optimization are chosen on the basis of experimental results, such as factor k and s in Interhand attraction and weights in the final optimization loss. In the future, we may set them as learnable parameters that can be automatically learned from data.

References

[1] 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. 1

[2] 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, 3, 4

[3] 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. 2

[4] 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*, 118:172–193, 2016. 3

[5] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 3