Supplemental Material for *Monocular 3D Reconstruction of Interacting Hands via Collision-Aware Factorized Refinements*

Yu Rong[1] Jingbo Wang[1] Ziwei Liu[2] Chen Change Loy[2]
[1]The Chinese University of Hong Kong [2]S-Lab, Nanyang Technological University
{ry017, wj020}@ie.cuhk.edu.hk {ccloy, ziwei.liu}@ie.cuhk.edu.hk

Table 1: **Weights and step size used in IHMR-OPT.** γ is the step size in optimizing the parameters. Objectives including $\mathcal{F}_{col}$ *etc*. are defined in Equ.(10) of the main paper.

Param $\hat{\boldsymbol{\rho}}\downarrow$	γ	$\mathcal{F}_{col}$	$\mathcal{F}_{2D}$	$\mathcal{F}_{3D}$	$\mathcal{F}_{\tau}$	$\mathcal{F}_{reg}$	$\mathcal{F}_{f}$
$\hat{\boldsymbol{\tau}}$	1e−4	1e−1	1e+1	1e+3	1e+3	1e−1	0.0
$\hat{\boldsymbol{\phi}}$	1e−2	1e−0	1e+1	1e+3	1e+3	1e−1	0.0
$\hat{\boldsymbol{\theta}}$	1e−2	1e−0	1e+1	1e+3	1e+3	1e−1	1e+5
$\hat{\boldsymbol{\beta}}$	1e−2	1e−0	1e+1	1e+3	1e+3	1e−1	0.0

1. Implementation Details

In this section, we introduce the implementation details of IHMR. We first introduce data preprocessing. Then we introduce the architecture and training details of the baseline CNN model used in stage-I. After that, we introduce the design details of the optimization. In the end, we introduce the architecture and training details of the MLP-based implementation.

1.1. Data Preprocessing

We tightly crop the input images surrounding the bounding boxes of the interacting hands. The hand bounding boxes are obtained from the 2D keypoint annotations. For InterHand2.6M dataset [7], we assume the known of hand types, *i.e*. single hand (left or right) or interacting hands. To increase the model's generalization ability to in-the-wild scenarios, where no such labels are provided, we follow the practice of InterNet [7] to additionally predict the hand type (left, right, or interacting) from input images. The trained model achieved 96% prediction accuracy on the whole test set of InterHand2.6M.

1.2. Baseline CNN Model

The input images are cropped, padded, and resized to 224×224. The encoder is a ResNet-50 [3]. The parameter prediction head is composed of three fully connected layers whose output dimensions are 1024, 1024, and 122. The 122 dimensions are composed of camera parameters ($\boldsymbol{\pi} \in \mathbb{R}^3$), shape parameters ($\boldsymbol{\beta} = (\boldsymbol{\beta}_l, \boldsymbol{\beta}_r) \in \mathbb{R}^{20}$), hand orientation ($\boldsymbol{\phi} = (\boldsymbol{\phi}_l, \boldsymbol{\phi}_r) \in \mathbb{R}^6$) finger pose parameters ($\boldsymbol{\theta} = (\boldsymbol{\theta}_l, \boldsymbol{\theta}_r) \in \mathbb{R}^{90}$), and right-to-left-hand relative translation ($\boldsymbol{\tau} \in \mathbb{R}^3$). The joint head follows the design as InterNet [7]. The whole framework, including the encoder, the parameter prediction head, and the joint head is trained end-to-end using Adam optimizer [4] with a learning rate 1e−3. The whole model converges after 20 epochs.

1.3. Optimization-based Implementation

The optimization-based method uses Adam optimizer [4] to directly update the estimated parameters $\hat{\boldsymbol{\rho}} \in \{\hat{\boldsymbol{\tau}}, \hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}\}$. When refining each parameter set, we use different step size γ and different weights for each objective listed in Equ.(8) and of the main paper. The values of step size and objective weights are listed in Tab. 1. We set smaller objective weights and step sizes in optimizing right-to-left-hand relative translation $\hat{\boldsymbol{\tau}}$. Otherwise, the optimization tends to unreasonably increase the overall distances between collided hands and thus generate worse results.

1.4. MLP-based Implementation

In the MLP-based implementation, each refinement stage is composed of a Multilayer Perceptron with four fully connected layers, whose output dimensions are 512, 256, 128, and K, where K is the dimensions of the corresponding parameters. The dimensions for each parameter are 3 for hand translation $\hat{\boldsymbol{\tau}}$, 6 for hand orientation $\hat{\boldsymbol{\phi}}$, 20 for shape parameters $\hat{\boldsymbol{\beta}}$ and 90 for finger poses $\bar{\boldsymbol{\theta}}$. The training set of these MLPs are composed of samples with closely interacting hands, following the same selecting criteria of "IH26M-Inter-Close" defined in Sec.4.1 of the main paper. We use the IHMR-OPT to obtain pseudo-ground-truth MANO parameters for train samples without MANO annotations. The loss weights for training the MLPs are the same as Equ.(5) of the main paper. Each MLP is trained for 2 epochs with the learning rate set as 1e−4. The optimizer is Adam [4].

1.5. HybrIK Implementation

To fill the performance gap between InterNet [7], we adopt a strong baseline following HybrIK [5]. We use the

Table 2: **Comparison with SOTA methods on InterHand2.6M using Vertex-Based Method.**. We additionally evaluate on subsets of InterHand2.6M [7] using vertex-based metrics MPVPE and I-MPVPE. AVE-P and MAX-P are adopted to estimate the collision status of the generated interacting hands.

Dataset & Metrics →	IH26M	IH26M-Inter			IH26M-Inter-Close		
Methods ↓	MPJPE / MPVPE ↓	MPJPE / MPVPE ↓	I-MPJPE / I-MPVPE ↓	AVE-P / MAX-P ↓	MPJPE / MPVPE ↓	I-MPJPE / I-MPVPE ↓	AVE-P / MAX-P ↓
Bouk *et al*. [1]	21.96 / 18.88	22.55 / 18.86	-	-	21.20 / 18.81	-	-
Pose2Mesh [2]	21.76 / 18.61	22.73 / 18.91	-	-	20.82 / 18.37	-	-
BiHand [9]	19.90 / 17.17	21.18 / 17.32	-	-	19.80 / 17.47	-	-
IHMR-Baseline	21.67 / 17.54	22.60 / 17.62	24.38 / 25.08	0.45 / 9.98	21.24 / 17.56	18.25 / 18.60	0.84 / 14.40
IHMR-MLP	22.79 / 18.16	23.37 / 17.82	21.26 / 21.85	0.33 / 8.03	21.55 / 17.58	16.60 / 16.82	0.68 / 12.67
IHMR-OPT	19.04 / 16.94	24.09 / 18.82	16.82 / 17.23	**0.13 / 3.75**	19.04 / 16.94	15.40 / 15.33	**0.33 / 7.30**
IHMR-Baseline*	17.05 / 17.18	17.54 / 16.71	14.44 / **12.45**	0.29 / 5.76	16.91 / 16.53	13.16 / 11.47	0.69 / 10.72
IHMR-MLP*	15.68 / **14.57**	**16.45 / 14.69**	13.64 / 11.81	0.26 / 5.34	15.76 / **14.56**	12.46 / **10.77**	0.61 / 10.07
IHMR-OPT*	**15.47** / 17.17	16.52 / 16.44	**13.49** / 13.01	0.23 / 4.48	**15.32** / 14.73	**11.90** / 11.23	0.34 / 7.48

Table 3: **Role of different optimization objectives.** We study the role of three objectives used in the optimization-based factorized refinement, namely 3D objectives $\mathcal{F}_{3D}$, 2D objectives $\mathcal{F}_{2D}$ and objective on hand translation $\mathcal{F}_{\tau}$.

$\mathcal{F}_{2D}$	$\mathcal{F}_{3D}$	$\mathcal{F}_{\tau}$	I-MPJPE ↓	AVE-P ↓
-	-	-	20.29	0.70
✓			18.99	0.14
	✓		17.07	0.22
		✓	36.76	**0.01**
✓	✓		**16.06**	0.27
✓		✓	17.89	0.12
	✓	✓	16.99	0.21
✓	✓	✓	16.94	0.20

Table 4: **Influence of the quality of pseudo-ground-truth 3D joints.** We add noise to ground-truth 3D joints and use these joints to serve as the pseudo 3D joints, namely $\dot{\boldsymbol{J}}_{3D}$. For still joints, we still use the original pseduo-ground-truth 2D joints, $\dot{\boldsymbol{J}}_{2D}$.

STD of Noise (mm) ↓	I-MPJPE ↓		AVE-P ↓	
	OPT	MLP	OPT	MLP
0	15.13	17.91	0.183	0.525
10	15.27	18.16	0.152	0.515
20	15.56	18.37	0.170	0.503
30	16.10	18.96	0.178	0.527
40	16.72	19.14	0.181	0.495

estimated 3D joints $\dot{\boldsymbol{J}}_{3D}$ from the joint head of Stage-I. To be specific, the relative right-to-left-hand translation $\hat{\boldsymbol{\tau}}$ is directly set as the subtraction of predicted left wrist joints by the right one. The global hand orientation $\hat{\phi}$ is calculated using the locations of wrists and the five palm joints through Singular Value Decomposition (SVD). The finger poses $\hat{\boldsymbol{\theta}}$ are calculated in the standard way of HybrIK. Please be noted that finger rotations are only composed of swing rotations. Therefore, the finger rotations can be directly solved from finger joint locations. In the last, we use the original shape parameters $\hat{\boldsymbol{\beta}}$ predicted from the baseline CNN model of Stage-I. We suggest to read the original paper of HybrIK [5] to have a better understanding of how HybrIK is adopted in our scenario.

2. More Quantitative Results.

Part of the training and testing data have pseudo-ground-truth MANO [8] parameters obtained from NeuralAnnot [6]. The number of samples with pseudo GT MANO annotations are 240K, 65K, 12K, and 4.5K for all four subsets. For samples with pseduo GT MANO annotations, We further calculate Mean Per Vertex Position Error (MPVPE) and I-MPVPE to reveal the quality of estimated joint rotations and shapes. They follow a similar definition as MPJPE and I-MPJPE. The results are included in Tab. 2. The conclusion we draw from Tab. 2 is similar to the conclusion we draw from Tab.1 of the main paper. The baseline models have similar performances. On the most challenging IH26M-Inter-Close test set, IHMR-OPT reduces the AVE-P by 60.7% while improving the accuracy of interacting pose estimation and 3D hand reconstruction by 14.3% and 16.0%. On the other hand, IHMR-MLP can reduce the collision, 3D finger poses error and 3D hand reconstruction error by 23.5%, 9.0%, and 9.7%.

3. More Ablation Studies

Following the same practice of the main paper, models in this subsection are evaluated on IH26M-Close-Inter, using I-MPJPE and AVE-P as the metrics. More ablation studies are included in the supplemental.

3.1. Influence of Different Optimization Objectives.

In this subsection, we study the role of several objectives used in the optimization implementation of the factorized refinement. The studied objectives include the 2D objective $\mathcal{F}_{2D}$, the 3D objective $\mathcal{F}_{3D}$ and the translation objective $\mathcal{F}_{\tau}$. The results are listed in Tab. 3. When 2D or 3D joint objectives are adopted, both the 3D joint estimation error and collision status can be reduced. When there is only the translation objective been adopted, although the collisions are almost totally removed, the joint estimation is ruined. In general, adopting all three objectives can lead to be best result with both good joint pose estimation and less collision status.

3.2. Influence of Joint Quality.

To evaluate the convergence of the proposed factorized refinement, we conduct experiments in which pseudo-ground-truth 3D joints $\dot{\boldsymbol{J}}_{3D}$ are replaced with ground-truth 3D joints with noise. The noise is sampled from Gaussian distribution with 0 mm mean and standard deviation ranging from 0 mm to 40 mm. For 2D joints, we still use the original pseudo-ground-truth 2D joints $\dot{\boldsymbol{J}}_{2D}$ obtained from the joint head. The results are listed in Tab. It is revealed that: (1) The proposed factorized refinement is robust to the noise. When the standard deviation of the added noise is increased to 40 mm, IHMR-OPT and IHMR-IHMR can still decrease the 3D joint estimation error by 17.6% and 5.7%. (2) The accuracy of 3D joint estimation is in proportion to the preciseness of the pseudo ground-truth 3D joints $\dot{\boldsymbol{J}}_{3D}$. (3) The effectiveness of collision removal has less correlation with the quality of $\dot{\boldsymbol{J}}_{3D}$.

4. More Qualitative Results

In this section, we show more qualitative results, including qualitative comparison between IHMR-MLP and IHMR-OPT, comparison between IHMR-MLP between single hand baseline, and quanlitative results demonstrating the effectiveness of finger regularization $\mathcal{F}_f$, and typical failure cases.

4.1. Comparison between Optimization and MLP

We show quanlitative comparison between IHMR-OPT and IHMR-MLP in Fig. 1. It is revealed that the optimization-based implementation can produce better 3D reconstructed hands with more precise joint estimation and fewer collisions.

4.2. Compare with Single-hand Methods

In this subsection, we qualitatively compare IHMRwith the single-hand baseline, *i.e*. Bouk *et al*. [1]. The results are shown in Fig. 2. It is revealed that the proposed IHMR can generate more precise interacting 3D hands than single-hand methods, which can only treat interacting hands as isolated single hands.

4.3. Influence of Finger Regularization

In this subsection, we show the influence of applying finger regularization $\mathcal{F}_f$ defined in Equ.(12) of the main paper. Several examples are listed in Fig. 3. It is revealed that without applying $\mathcal{F}_f$, optimization methods tend to generate twisted fingers as marked by red circles in Fig. 3.

4.4. Failure Cases.

We show several typical failure cases in Fig. 4. As the results show, typical failure cases are caused by challenging poses and occlusions.

References

[1] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. ii, iii, iv

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision*, 2020. ii

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. i

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015. i

[5] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. i, ii

[6] Gyeongsik Moon and Kyoung Mu Lee. Neuralannot: Neural annotator for in-the-wild expressive 3d human pose and mesh training sets. *arXiv preprint arXiv:2011.11232*, 2020. ii

[7] 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*, 2020. i, ii

[8] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 2017. ii

[9] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In *British Machine Vision Conference*, 2020. ii

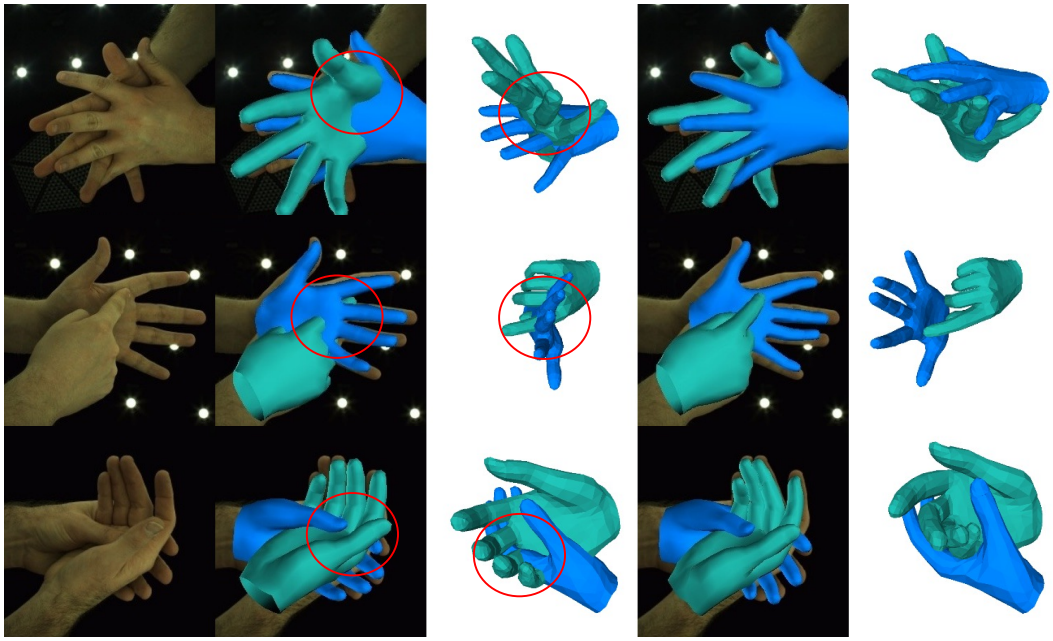

Figure 1: **Qualitative Comparison between IHMR-MLP and IHMR-OPT.** We qualitatively compare between IHMR-MLP and IHMR-OPT. Collided regions are marked with red circles.

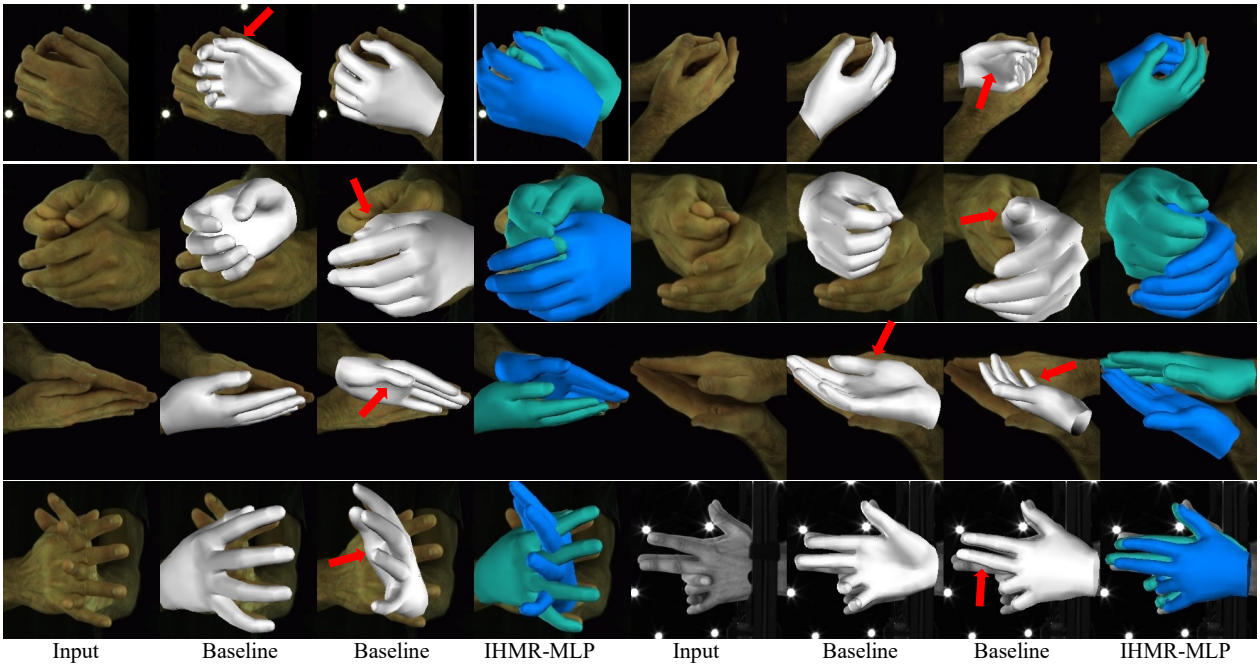

Figure 2: **Compare with SOTA single-hand method.** In this figure, we qualitatively compare with single-hand baseline, *i.e.* Bouk [1]. From left to right are input images, predicted right hands from the baseline, predicted left hands from the baseline and predicted interacting hands from IHMR-MLP. Imprecise hand poses or finger poses generated by the single-hand method are marked with red arrows.

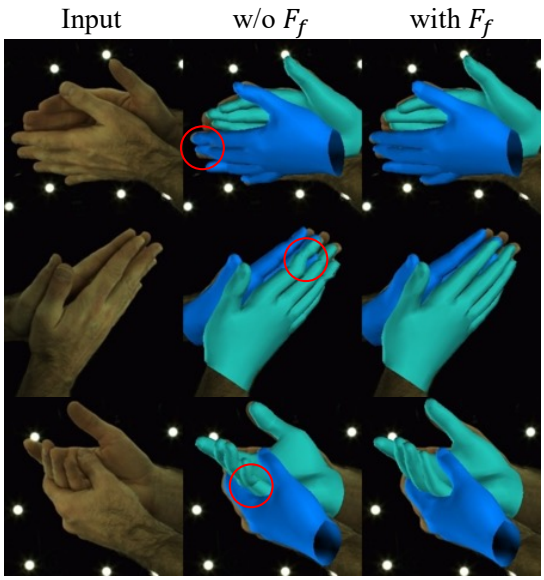

Figure 3: **Role of $\mathcal{F}_f$.** In this figure, we demonstrate the role of $\mathcal{F}_f$ plays in optimization. From left to right are input images, optimized results without using $\mathcal{F}_f$, optimized results with using $\mathcal{F}_f$. Twisted fingers generated by optimization without using finger regularization $\mathcal{F}_f$ are marked with red circles.

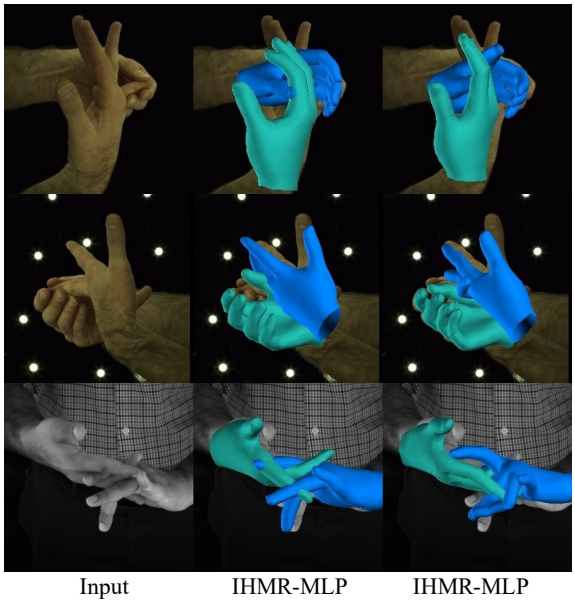

Figure 4: **Typical Failure Cases.** Typical failure cases of IHM-Rare caused by challenging finger poses and collusions.