

Monocular 3D Reconstruction of Interacting Hands via Collision-Aware Factorized Refinements

Yu Rong¹ Jingbo Wang¹ Ziwei Liu² Chen Change Loy²

¹The Chinese University of Hong Kong ²S-Lab, Nanyang Technological University

{ry017, wj020}@ie.cuhk.edu.hk {ziwei.liu, ccloy}@ntu.edu.sg

Abstract

3D interacting hand reconstruction is essential to facilitate human-machine interaction and human behaviors understanding. Previous works in this field either rely on auxiliary inputs such as depth images or they can only handle a single hand if monocular single RGB images are used. Single-hand methods tend to generate collided hand meshes, when applied to closely interacting hands, since they cannot model the interactions between two hands explicitly. In this paper, we make the first attempt to reconstruct 3D interacting hands from monocular single RGB images. Our method can generate 3D hand meshes with both precise 3D poses and minimal collisions. This is made possible via a two-stage framework. Specifically, the first stage adopts a convolutional neural network to generate coarse predictions that tolerate collisions but encourage pose-accurate hand meshes. The second stage progressively ameliorates the collisions through a series of factorized refinements while retaining the preciseness of 3D poses. We carefully investigate potential implementations for the factorized refinement, considering the trade-off between efficiency and accuracy. Extensive quantitative and qualitative results on large-scale datasets such as InterHand2.6M demonstrate the effectiveness of the proposed approach.

1. Introduction

Capturing 3D interacting hand motion from monocular single RGB images can facilitate numerous downstream tasks including AR/VR [8, 46] and social signal understanding [12, 25]. Previous works on motion capture of two interacting hands [2, 24, 27, 38, 42, 45] mainly rely on depth images, multi-view images or image sequences as input. These methods cannot be easily applied to monocular RGB images. Recently, there is a surge of interest to capture 3D single hand motion from monocular RGB images [1, 3, 7, 17, 47, 51]. For example, MANO [29], a 3D hand mesh model parameterized by pose parameters and shape parameters, is proposed to represent 3D hands. Deep neural networks are leveraged to regress the model parameters

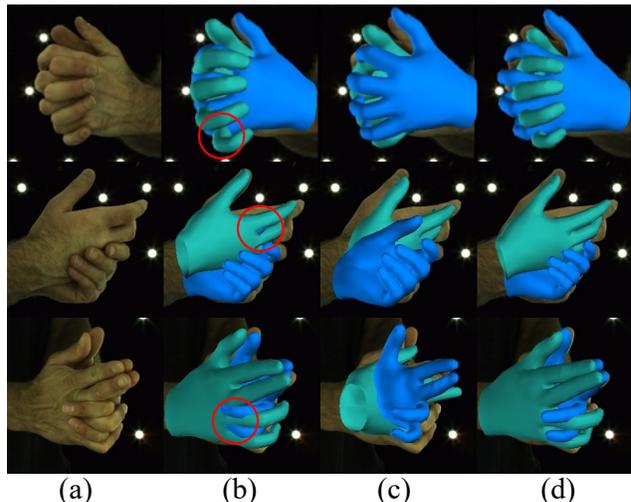


Figure 1: **Interacting 3D Hands Reconstruction.** (a) Input RGB images with closely interacting hands. (b) A baseline model follows the single-hand based method [3] to reconstruct two hands. The method tends to generate collided 3D hands. Collided regions are marked with red circles. (c) Results obtained by directly adding a collision-aware loss [11] in training. (d) Results produced by the proposed method with mitigated collisions.

or coordinates of the mesh vertices. Although these methods achieve good performance in single-hand scenarios, it remains challenging to estimate interacting hands given just monocular RGB images. Typical results of a baseline model adapted from the single-hand methods [3] are shown in Fig. 1 (b). One major obstacle is that these methods do not explicitly model the relationship between two closely interacting hands, and therefore, inevitably generate unnatural results with hands colliding with each other.

An intuitive idea to diminish collisions is by applying collision-aware losses, inspired by previous works that model body-centric interactions [9, 11, 50]. These losses are useful to increase the overall distances between interacting subjects, the main cause of collisions found across interacting bodies. Collisions observed between interacting hands are attributed to multiple factors, such as imprecise

hand shapes or finger poses. Simply adding collision losses in model training inevitably makes a model converge to a trivial solution that merely pulls two hands away. While collisions are dismissed, the accuracy of 3D hand estimation is also compromised. We illustrate several examples in Fig. 1 (c). A similar phenomenon has also been observed by previous body-motion studies, *e.g.*, Jiang *et al.* [11].

In this work, we introduce a two-stage framework that estimates 3D hand poses and shapes of two closely interacting hands with precise 3D poses and little collisions. In the first stage, we adapt a convolutional neural network (CNN) with a similar architecture as single hand methods [3, 51] to predict initial hand meshes of two hands. Specifically, the CNN takes a single RGB image as input and regresses two-hand parameters as its output. In this stage, collision-aware losses are not adopted – we deliberately allow the model to produce collided output but focus on generating acceptable and accurate pose results. These initial predictions provide good initialization for the second stage. Several examples are shown in Fig. 1 (b). From the coarse results obtained in the first stage, we observe that collisions of different samples are caused by distinct factors. For example, the collisions shown in the first row of Fig. 1 (b) are caused by inaccurate hand shape estimations while imprecise relative hand location causes the collisions in the second row. In the third row, collisions are mainly caused by the imperfect finger pose estimation. An effective way to improve each collided initial prediction is to focus on its specific root cause of error instead of jointly optimizing for all components.

To this end, we present a novel factorized refinement strategy in the second stage to ameliorate the collision issue. We decompose the causing factor of error and correct one factor at a time. In this work, we consider different approaches to perform the factorized refinement, *i.e.*, through the conventional gradient-descent based optimization and a more effective alternative that applies MLP as a proxy for optimization. Using the proposed factorized refinement strategy, the collision issues can be substantially mitigated. We show several recovered 3D hand meshes using the proposed two-stage framework in Fig. 1 (d). Compared with baseline results shown in Fig. 1 (b) or results from a model trained with naively applying collision-aware losses, as shown in Fig. 1 (c), our method can generate precise 3D hand meshes with modest collisions.

Our contributions can be summarized as follows: 1) We introduce the first 3D interacting hand motion capture framework that only needs monocular single RGB images as inputs, without any other auxiliary sensory inputs. 2) We comprehensively investigate the collision issue that exists in interacting hands motion capture, and subsequently design a simple yet effective strategy to resolve the problem. 3) We present two alternatives for factorized refinement that carefully consider the trade-off in accuracy and

speed. We perform extensive evaluations on the proposed method on large-scale interacting hand datasets, *e.g.*, InterHand2.6M [22]. In comparison to existing methods, the proposed approach achieves 71.4% reduction in the generated collisions and improves pose estimation by 16.5%.

2. Related Work

3D Single Hand Pose Estimation. Early studies [26, 35, 36, 40] take depth images as input and use optimization or discriminative methods to predict 3D joint locations. Subsequent works [6, 20, 41, 44, 49] replace the time-consuming optimization with deep neural networks. Follow-up works [4, 23, 32, 34, 48, 52] start to take RGB images as inputs and deploy CNN to estimate the 3D hand poses. Inspired by recent progress in 3D body mesh recovery [13, 15, 16, 30, 31, 32], more recent approaches [1, 3, 7, 17, 47, 51] try to estimate 3D single-hand meshes from monocular RGB images. Most works [1, 3, 47, 51] use a CNN to predict the parameters of the MANO model [29]. There are also other works directly regress the 3D vertices of the hand. Ge *et al.* [7] and Pose2Mesh [5] use graph neural networks. Kulon *et al.* [17] use mesh convolution. Lin *et al.* [19] use Transformers [43] as the backbone. Although these methods perform well on single hands, it is non-trivial to adapt them to cope with interacting hands. Single-hand methods do not explicitly model the interrelationship between hands and thus will inevitably generate erroneous hand meshes, *e.g.* hands colliding with each other.

3D Interacting Hand Pose Estimation. Oikonomidis *et al.* [27] firstly use Particle Swarm Optimization (PSO) to fit the 3D interacting hand motion that aligns best with the RGB-D inputs. Ballan *et al.* [2] solve this problem by estimating both finger-salient points and overall hand poses. They leverage the multi-view RGB sequences as the input. Tzionas *et al.* [42] extend the idea of discriminative salient points by introducing physics simulation. Taylor *et al.* [38] introduce a new implicit hand model to facilitate simultaneous optimization over both hand poses and hand surfaces. Smith *et al.* [33] use an elastic volume deformation and a collision response model to recover dense hand surface from multi-view sequences. More recent works of Muller *et al.* [24] and Wang *et al.* [45] share the similar pipeline. Specifically, deep neural networks are firstly used to predict the dense correspondence between pixels and MANO [29] surface. MANO parameters are then obtained by applying optimization to the obtained intermediate predictions. Muller *et al.* [24] use single-view depth sequences while Wang *et al.* [45] use single-view RGB sequence as input. Limited by the different input modality, these specialized methods cannot be directly applied to single RGB images. InterNet [22], although can cope with monocular RGB images, only produces 3D hand joint positions, which only have limited applications. To fill the gap, we propose a

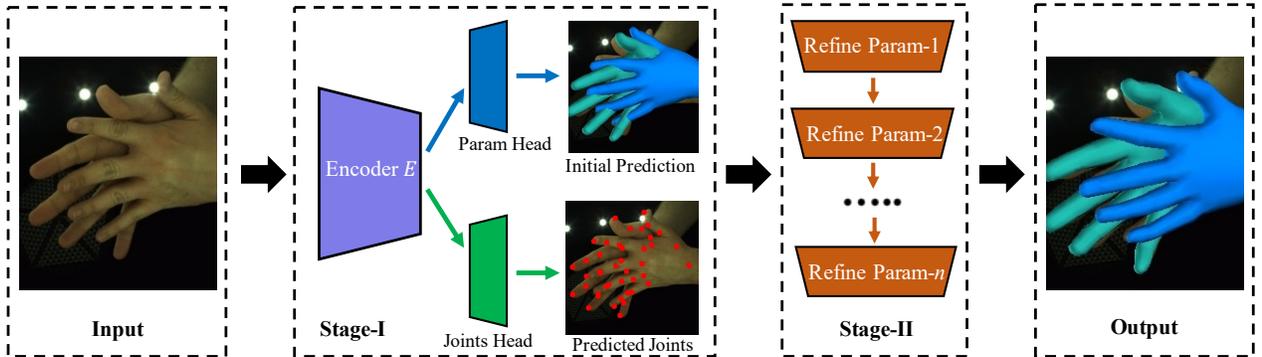


Figure 2: **Framework overview.** The proposed IHMR takes monocular RGB images as input and generates the 3D meshes represented in MANO [29]. The framework has two stages. The first stage adopts a convolutional neural network composed of an encoder, a MANO parameter prediction head, and a 2D/3D joint prediction head. It outputs the initial predicted parameters $\hat{\Theta}$ and predicted 2D and 3D joints, namely \hat{J}_{2D} and \hat{J}_{3D} . By taking initial predictions as input, the second stage performs factorized refinement to gradually remove the collisions from the first stage and produce refined predictions. The detailed architecture of each refinement module is depicted in Fig. 3.

new framework to simultaneously estimate both poses and shapes of interacting hands from single RGB images.

Collision Detection across Interacting Hands. Oikonomidis *et al.* [27] considers adjacent fingers interpenetrate by calculating abduction-adduction difference. Ballan *et al.* [2] and Tzionas *et al.* [42] use bounding volume hierarchies (BVH) [39] to detect each collided faces. Muller *et al.* [24] and Wang *et al.* [45] use 3D Gaussians to model interpenetrate between fingers and palms. Smith *et al.* [33] does not only detect collided meshes but also adopt elasticity physical models to simulate the soft tissue of the hand skins. In this paper, we use a modified Signed Distance Function (SDF) [9, 11] which can model the penetration relations more precisely than finger-level [27], BVH [2, 42] or 3D Gaussians [24, 45]. Besides, we propose to use the factorized refinement to effectively diminish the collision while retaining the accuracy of the estimated finger poses.

3. Methodology

To reconstruct closely interacting 3D hands from monocular single RGB images, we propose a two-stage framework called Interacting Hand Mesh Recovery (IHMR). As depicted in Fig. 2, in the first stage, a CNN similar to a previous single-hand model [3] is adapted to generate initial predictions of two hands from input images. No collision-aware losses are applied in this stage so that the model can focus on generating accurate hand poses. A factorized refinement strategy is introduced in the second stage to gradually diminish the collisions observed in the first stage while retaining the accuracy of the estimated 3D hands. We provide two different implementations for the proposed strategy, namely a optimization-based method with better accuracy and a neural-network based method with higher speed. Next, we first introduce the 3D hand model used in this paper. Then we discuss the design of the CNN baseline model

and the factorized refinement strategy. Lastly, we present the details of two different implementations of the latter.

3.1. 3D Hand Model

We use MANO [29], a 3D hand mesh model parameterized by the shape and pose parameters. The shape parameters $\beta_h \in \mathbb{R}^{10}$ control the overall shape of hands. The pose parameters are composed of two parts. The first part $\phi_h \in \mathbb{R}^3$ controls the global hand orientation. The second part $\theta_h \in \mathbb{R}^{3 \times 15}$ represents the 3D rotations of each finger joint, relative to their parents on the predefined kinematics skeleton. Hand orientation and finger poses are both represented in axis-angle format. Given shape parameters β_h , hand orientation ϕ_h and finger poses θ_h , MANO model W_h calculates the 3D vertices $V_h \in \mathbb{R}^{3 \times 778}$ as:

$$V_h = W_h(\beta_h, \phi_h, \theta_h). \quad (1)$$

The original MANO model only supports single hands. We extend it to support interacting hand scenarios by using two sets of parameters, namely $(\beta_l, \phi_l, \theta_l)$ for left hands and $(\beta_r, \phi_r, \theta_r)$ for right hands. Another parameter $\tau \in \mathbb{R}^3$ is added to represent the relative translation from the right hand to the left one. For simplicity of notation, we stack the same parameters of left and right hands together as $\beta = (\beta_l, \beta_r)$, $\phi = (\phi_l, \phi_r)$, and $\theta = (\theta_l, \theta_r)$. The whole process of obtaining 3D interacting hand vertices from aforementioned parameters is formulated as:

$$V = W(\beta, \phi, \theta, \tau), \quad (2)$$

$$= (W_{lh}(\beta_l, \phi_l, \theta_l) + \tau, W_{rh}(\beta_r, \phi_r, \theta_r)),$$

where W_{lh} and W_{rh} represent MANO models for left and right hands. $V = (V_l, V_r)$ is built from stacking left and right hand vertices. 3D joints $J_{3D} \in \mathbb{R}^{3 \times 42}$ are obtained simultaneously in the process of linear blending skin.

3.2. Basic 3D Hand Reconstruction Model

Architecture. The baseline CNN model is composed of an encoder, a MANO parameter prediction head and a 2D/3D joint prediction head. Given single images as input, the encoder produces the encoded features \mathbf{f} , which are then fed to the parameter prediction head. The parameter prediction head outputs the estimated parameters $\hat{\Theta} = (\hat{\beta}, \hat{\phi}, \hat{\theta}, \hat{\tau})$, as defined in Eq. (2). Besides these parameters, the parameter head also predicts a set of weak-perspective camera parameters $\pi \in \mathbb{R}^3$ to project the 3D joints into 2D space. The camera parameters $\pi = (s, \mathbf{t})$ is composed of a scale factor s and 2D translation $\mathbf{t} \in \mathbb{R}^2$. Given camera parameters π , the 2D joints \mathbf{J}_{2D} can be obtained as $\mathbf{J}_{2D} = s\Pi(\mathbf{J}_{3D}) + \mathbf{t}$, where Π is the orthogonal projection.

3D Losses. To train the baseline model, we adopt a set of 3D losses that include MANO parameter losses and 3D joint loss. These losses are defined as follows:

$$\begin{aligned} L_{\Theta_h} &= \|\hat{\beta} - \hat{\beta}\|_2^2 + \|R(\hat{\theta}) - R(\hat{\theta})\|_2^2, \\ L_{J_{3D}} &= \|\hat{\mathbf{J}}_{3D} - \hat{\mathbf{J}}_{3D}\|_2^2, \\ L_{\tau} &= \|\hat{\tau} - \hat{\tau}\|_2^2, \\ L_{reg} &= \|\hat{\beta}_l - \hat{\beta}_r\|_2^2, \end{aligned} \quad (3)$$

where $\hat{\beta}, \hat{\tau}, \hat{\theta}, \hat{\mathbf{J}}_{3D}$ represent ground-truth MANO shape parameters, relative hand translation, finger poses and 3D joints. $\hat{\tau}$ is obtained from ground-truth 3D locations of left and right hand wrists. R is the Rodrigue’s rotation formula, which converts the finger poses from the axis-angle format to the rotation-matrix format. We follow this practice of previous works [13] for better numerical stability. L_{reg} is used to constrain the shape variation between two hands belonging to the same person. We also tried surface-based 3D losses such as vertex coordinate loss, surface normal and edge losses used by Pose2Mesh [5], but empirical results show that these losses bring no performance gain.

2D Losses. We also adopt a 2D loss defined as:

$$L_{J_{2D}} = \|\mu \cdot (\hat{\mathbf{J}}_{2D} - \hat{\mathbf{J}}_{2D})\|_1, \quad (4)$$

where μ is the visibility indicator and $\hat{\mathbf{J}}_{2D}$ is the ground-truth 2D joints.

Overall Loss. The overall loss L is defined as follows:

$$L = \lambda_1 L_{\Theta_h} + \lambda_2 L_{\tau} + \lambda_3 L_{J_{3D}} + \lambda_4 L_{reg} + \lambda_5 L_{J_{2D}} \quad (5)$$

In our experiments, the values of these λ s are set to be [10.0, 10.0, 10.0, 0.1, 10.0], respectively.

3.3. Collision-Aware Factorized Refinements

3.3.1 Collision-aware Loss

In this second stage, we adopt the penetration loss used by previous work [9, 11] to reduce the collisions observed

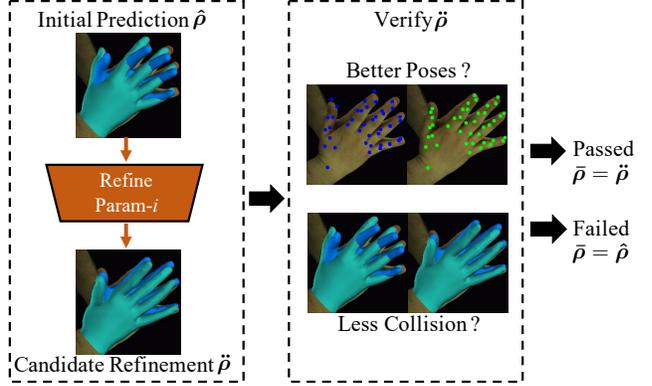


Figure 3: **Refinement with verification.** Given original predictions generated by the baseline CNN model or previous refinement stages, the refinement module of current stage aims at refining the corresponding parameters and generating the candidate refinement. Some criteria are used to decide whether to keep the original predictions before refinement or adopt the refined parameters.

in the first stage. Following their practice, we first use a modified Signed Distance Field (SDF) for each hand mesh, which is defined as follows:

$$\psi_h(x, y, z) = -\min(\text{SDF}(x, y, z), 0), \quad (6)$$

For points (x, y, z) inside the hand meshes, the $\psi_h(x, y, z)$ takes positive values, which are in proportion to the points’ distances to the hand surfaces. The values take zero for outside points. The SDF is obtained via voxelizing the original 3D hand meshes to 3D grids with dimension $N_p \times N_p \times N_p$. We compute ψ_l for left hands and sample the SDF value of each vertex of right hands, and vice versa. The collision loss is calculated as follows:

$$L_{col} = \sum_{v \in V_l} \psi_r(v) + \sum_{v \in V_r} \psi_l(v), \quad (7)$$

where ψ_l and ψ_r are SDF functions of left and right hand meshes, while V_l and V_r are left and right hands vertices.

Recall that we do not add the collision loss L_{col} in training the baseline CNN model of the first stage. It is because adding collision-aware losses in model training will compromise the estimated 3D hands, as depicted in Fig. 1 (c). Besides, even with a rigorous adjustment on the loss weight of L_{col} , we found that the model cannot converge to provide a balanced output, where pose estimation and collision status are both satisfactory. We verify this phenomenon in our experiments (Fig. 6), in which we list the baseline model’s performances obtained by setting different loss weights of L_{col} . In view of this, we propose to have factorized refinement as a second stage to refine the outputs generated by the first-stage CNN model.

3.3.2 Factorized Refinement

We divide the cause factors of inter-hand collisions into four different parameter sets, namely imprecise right-to-left-hand translation $\hat{\tau}$, imprecise hand orientations $\hat{\phi}$, imprecise hand shapes $\hat{\beta}$ and imprecise hand finger poses $\hat{\theta}$. The goal of factorized refinement is to refine each cause factor given each collide sample. However, due to the lack of corresponding annotations or explicit signals, it is challenging to use prior knowledge to precisely identify the major cause factors. Therefore, we adopt stagewise refinement with verification, as shown in Stage-II of Fig. 2, with each stage detailed in Fig. 3. The whole process of factorized refinement is summarized in Algorithm 1.

Given each collided sample, we try to refine all four parameters in a stage-wise manner, following the order of $\hat{\rho} \in \{\hat{\tau}, \hat{\phi}, \hat{\theta}, \hat{\beta}\}$. The order is insignificant as we observed from our experiments. We denote the initial predictions as $\hat{\rho}$, the candidate obtained after the refinement module as $\check{\rho}$, and the refined parameters as $\bar{\rho}$. In each stage, we first refine one of the parameter sets, $\hat{\rho}$, and obtain a candidate parameter set $\check{\rho}$. Then we check whether the candidate set could lead to better results, *e.g.*, hand meshes with less collisions and lower 3D joint errors (detailed in next paragraph). If the results are improved after verification, we use the candidate $\check{\rho}$ as the final refined parameters $\bar{\rho}$. Otherwise, we still keep the original values $\hat{\rho}$ as the output of this stage. Once we are done with the refinement of a parameter set we move to the next stage to refine another set of parameters.

For each refinement stage, we perform verification to decide if the candidate parameters could lead to a better performance. We adopt two criteria, *i.e.*, collision errors \mathcal{E}_{col} and 3D joint errors \mathcal{E}_{3D} . Collision errors can be calculated as Eq. (7) using the estimated vertices \hat{V} . To calculate the 3D joint errors during inference time, we add another head onto the baseline CNN model to directly predict the 3D joints \hat{J}_{3D} , and use them as the pseudo-ground-truth 3D joints to calculate the 3D joint error as $\mathcal{E}_{3D} = \|\hat{J}_{3D} - \check{J}_{3D}\|_2^2$. We follow the same architecture of the InterNet [37] to use 3D heatmaps to build this additional joint prediction head. More details about the joint head are included in the supplemental material. We use \check{J}_{3D} instead of regressed 3D joints from initial MANO predictions \hat{J}_{3D} because it is revealed by the previous works that 3D joints predicted in the heatmap format have better accuracy than the joints regressed from mesh models [18].

Optimization-based Implementation. Our first implementation of stage-wise refinement uses gradient-descent based optimization, which updates the parameters by minimizing the following objective function:

$$\mathcal{F}_{opt} = \mathcal{F}_{col} + \mathcal{F}_{2D} + \mathcal{F}_{3D} + \mathcal{F}_{\tau} + \mathcal{F}_{reg} + \mathcal{F}_f, \quad (8)$$

where \mathcal{F}_{col} is computed using Eq. (7). The term \mathcal{F}_{2D} is computed in the same way as Eq. (4) by replacing the

Algorithm 1 Factorized Refinement Strategy.

Require: Initial parameters $\hat{\Theta} = (\hat{\beta}, \hat{\phi}, \hat{\theta}, \hat{\tau})$.

Require: Pseudo-ground-truth 3D Joints \hat{J}_{3D} .

Require: Initialize refined parameters $\bar{\Theta}$ as $\hat{\Theta}$.

```

1: function OBTAINERROR( $\Theta, \hat{J}_{3D}$ )
2:    $V = W(\Theta)$ 
3:    $J_{3D} = (J \cdot V^T)^T$ 
4:    $\mathcal{E}_{col} = \Psi(V)$ 
5:    $\mathcal{E}_{3D} = \|J_{3D} - \hat{J}_{3D}\|_2^2$ 
6:   return  $\mathcal{E}_{col}, \mathcal{E}_{3D}$ 
7: end function
8:  $\bar{\mathcal{E}}_{col}, \bar{\mathcal{E}}_{3D} = \text{ObtainError}(\bar{\Theta}, \hat{J}_{3D})$ 
9: for  $\bar{\rho}$  in  $\{\hat{\beta}, \hat{\phi}, \hat{\theta}, \hat{\tau}\}$  do
10:   $\check{\rho} = \text{Refine}(\bar{\rho})$ 
11:   $\check{\Theta} = (\bar{\Theta} \setminus \bar{\rho}) \cup \{\check{\rho}\}$ 
12:   $\check{\mathcal{E}}_{col}, \check{\mathcal{E}}_{3D} = \text{ObtainError}(\check{\Theta}, \hat{J}_{3D})$ 
13:  if  $\check{\mathcal{E}}_{col} < \bar{\mathcal{E}}_{col}$  and  $\check{\mathcal{E}}_{3D} < \bar{\mathcal{E}}_{3D}$  then
14:     $\bar{\rho} = \check{\rho}$ 
15:     $\bar{\mathcal{E}}_{col} = \check{\mathcal{E}}_{col}$ 
16:     $\bar{\mathcal{E}}_{3D} = \check{\mathcal{E}}_{3D}$ 
17:  end if
18: end for

```

ground-truth 2D joints J_{2D} with \hat{J}_{2D} , which is also estimated by the additional joint head used to predict \hat{J}_{3D} . The term \mathcal{F}_{τ} is computed in a similar way as L_{τ} . The pseudo-ground-truth right-to-left-hand translation is obtained by subtracting the 3D left wrist location by the right one. The 3D joint locations are retrieved from \hat{J}_{3D} . The term \mathcal{F}_{reg} is the same as L_{reg} in Eq. (3).

The last term \mathcal{F}_f is used to avoid generating twisted fingers during optimization. We follow previous study [51] to formulate \mathcal{F}_f . Let p_a, p_b, p_c, p_d represent joints on the fingers from the tip to the palm. $\vec{V}_{ab} = p_b - p_a$ represents the finger vectors. There are two constraints listed as follows:

$$\begin{aligned} \mathcal{C}_1 &= (\vec{V}_{ab} \times \vec{V}_{bc}) \cdot \vec{V}_{cd} = 0 \\ \mathcal{C}_2 &= (\vec{V}_{ab} \times \vec{V}_{bc}) \cdot (\vec{V}_{bc} \times \vec{V}_{cd}) \geq 0 \end{aligned} \quad (9)$$

The finger error \mathcal{F}_f is defined as:

$$\mathcal{F}_f = \|\mathcal{C}_1\| + \min(\mathcal{C}_2, 0). \quad (10)$$

Detailed loss weights and learning rate used in each stage are listed in the supplemental. At the end of each stage, verification criteria \mathcal{E}_{col} and \mathcal{E}_{3D} are used to select the valid updated parameters that can improve the performances.

MLP-based Implementation. The optimization-based method offers a good performance but it runs at a slow speed (0.9 FPS). In view of this, we provide another implementation of factorized refinement using Multilayer Perceptrons (MLP), which can run in real time (29.6 FPS).

Specifically, a separate MLP is deployed to refine each parameter set. The input of MLPs is the concatenation of encoded features \mathbf{f} generated by image encoder E and the initial parameter predictions $(\hat{\beta}, \hat{\phi}, \hat{\theta}, \hat{\tau})$. The output is the candidate parameters $\check{\rho}$, which is subjected to the verification step discussed earlier. To train MLPs, we build a train subset consisting of only closely interacting samples. The training losses include 3D losses defined in Eq. (3), 2D losses defined in Eq. (4) and collision loss defined in Eq. (7). Similar to the optimization-based implementation, criteria including \mathcal{E}_{col} and \mathcal{E}_{3D} are adopted during both train and test phases to filter out the candidate parameters that cannot lead to improved results. Implementation details of the MLPs are included in the supplemental material.

4. Experiments

4.1. Experimental Settings

Datasets. We majorly conduct experiments on InterHand2.6M [22], which provides RGB images with annotated 2D and 3D joints. The dataset also provides pseudo-ground-truth MANO [29] parameters obtained from NeuralAnnot [21] for part of samples. Follow the practice of the original paper, we use the 5 FPS version of the released data. Besides, we neglect the multi-view information and treat all data as single-view samples. Our models are trained on the whole train set evaluated on the test set. To focus our evaluation on the interacting hands, we select interacting samples from the whole test set. Following the same definition as the original paper, we attribute samples with more than 30 valid ground-truth 3D joints as “interacting”. Besides the interacting subset, another subset composed of only **closely** interacting samples is further filtered out. Given each interacting sample, we first calculate the per-joint “inter-distance”, which is defined as the smallest Euclidean distance from this joint to the joints belonging to the other hand. Closely interacting samples are defined as the sample whose average inter-distance of all valid joints is smaller than 40 mm. We simplify the names of the original test set, the interacting subset, and the closely interacting subset as “IH26M-All”, “IH26M-Inter”, and “IH26M-Inter-Close”. We also train and evaluate our models on RGB2Hand dataset [45] and Tzionas dataset [42]. We follow the same practice of InterNet [22] to randomly split the origin Tzionas dataset [42] into train and test set at a ratio of 9:1 and additionally train and evaluate our models on the split sets.

Evaluation Metrics. We first adopt a single-hand style metric, 3D Mean Per Joint Position Error (MPJPE). In calculating MPJPE, the left and right hand joints are separately aligned to the ground-truth left and right wrists. We adopt another interacting-hand style metric I-MPJPE that synchronically aligns both the left and right hands joints to the ground-truth joints using Procrustes Analysis. Fol-

lowing the practice of RGB2Hands [45], we only normalize the scales and translations while leaving the estimated rotations unchanged. Furthermore, we adopt two more metrics to reveal the collision status of the reconstructed interacting hands. They are Average Penetration Depth (AVE-P) of each hand vertices penetrating into the other hand, and Maximum Penetration Depth (MAX-P) of all vertices belonging to a pair of hand. In the last, we evaluate samples with pseudo-ground-truth MANO annotations using Mean Per Vertex Position Error (MPVPE) and corresponding I-MPVPE. These results are listed in the supplemental. Units of the aforementioned metrics are all millimeters (mm).

4.2. Comparison

Comparison on InterHand2.6M. The quantitative results on InterHand2.6M [22] are listed in Tab. 1. We first compare performances of three single-hand methods (Bouk *et al.* [3], Pose2Mesh [5] and BiHand[47]) on test set of InterHand2.6M using single-hand metric MPJPE. All these three methods are finetuned on the train set of InterHand2.6M. As the models’ performances are close, we choose Bouk *et al.* [3] as the baseline model to build up IHMR for its simplicity.

Due to the lack of similar works, we majorly compare the variants of IHMR. They are the baseline CNN model (IHMR-Baseline), IHMR with factorized refinement implemented in MLP (IHMR-MLP), and IHMR implemented in optimization (IHMR-OPT). We also test the models’ inference speed on a single RTX 2080Ti GPU. The results show that the proposed factorized refinement strategy can effectively diminish the collisions between the interacting hands while producing more precise 3D poses. On the most challenging IH26M-Inter-Close, IHMR-OPT reduces the AVE-P by 71.4% while improving the interacting pose estimation accuracy by (in terms of I-MPJPE) by 16.5%. On the other hand, IHMR-MLP can reduce the collision and 3D finger poses errors by 26% and 10.4%. We believe the performance gap between IHMR-MLP and IHMR-OPT is caused by their inherent differences. Optimization is better at modeling high-frequency information such as per-vertex collisions. On the other hand, IHMR-MLP achieves real-time performance, which is more suitable for applications such as live streaming that demand real-time processing.

Furthermore, we adopt HybrIK [18] as an additional strong baseline to fill the gap between InterNet [22]. Implementation details are included in the supplemental. Based on this strong baseline, we further adopt factorized refinement. The results are listed in the last 4th to 2nd rows (marked with *) of Tab. 1. By following HybrIK, the I-MPJPE gap between InterNet and IHMR-Baseline* is reduced from 8.3 mm to 3.5 mm. Furthermore, the proposed factorized refinement can still effectively increase the accuracy of 3D joints estimation and reduce collisions despite

Table 1: **Comparison with SOTA methods on InterHand2.6M [22]**. We apply MPJPE and I-MPJPE to evaluate 3D joint estimation accuracy in the single-hand style and interacting-hand style. AVE-P and MAX-P are adopted to estimate the collision status of the generated interacting hands. We also test the speed of the variants of IHMR, the unit is frame-per-second (FPS).

Dataset & Metrics →	IH26M	IH26M-Inter			IH26M-Inter-Close			Speed
Methods ↓	MPJPE ↓	MPJPE ↓	I-MPJPE ↓	AVE-P / MAX-P ↓	MPJPE ↓	I-MPJPE ↓	AVE-P / MAX-P ↓	Speed (FPS) ↑
Bouk <i>et al.</i> [3]	27.14	31.46	-	-	26.64	-	-	-
Pose2Mesh [5]	27.10	32.11	-	-	25.97	-	-	-
BiHand [47]	25.10	28.23	-	-	24.16	-	-	-
IHMR-Baseline	25.45	29.70	39.38	0.47 / 10.65	26.30	20.29	0.70 / 14.10	76.8
IHMR-MLP	24.51	28.72	32.75	0.32 / 8.34	24.74	18.18	0.52 / 12.00	29.6
IHMR-OPT	22.88	29.57	23.82	0.09 / 3.15	24.61	16.94	0.20 / 6.23	0.89
IHMR-Baseline*	18.52	21.78	20.03	0.27 / 5.53	19.63	15.35	0.64 / 10.35	75.2
IHMR-MLP*	17.61	21.23	19.45	0.25 / 5.34	18.60	14.42	0.44 / 9.55	29.6
IHMR-OPT*	17.12	20.66	19.05	0.21 / 4.48	17.66	13.56	0.27 / 6.27	0.89
InterNet [22]	14.21	18.04	17.36	-	16.19	11.91	-	-

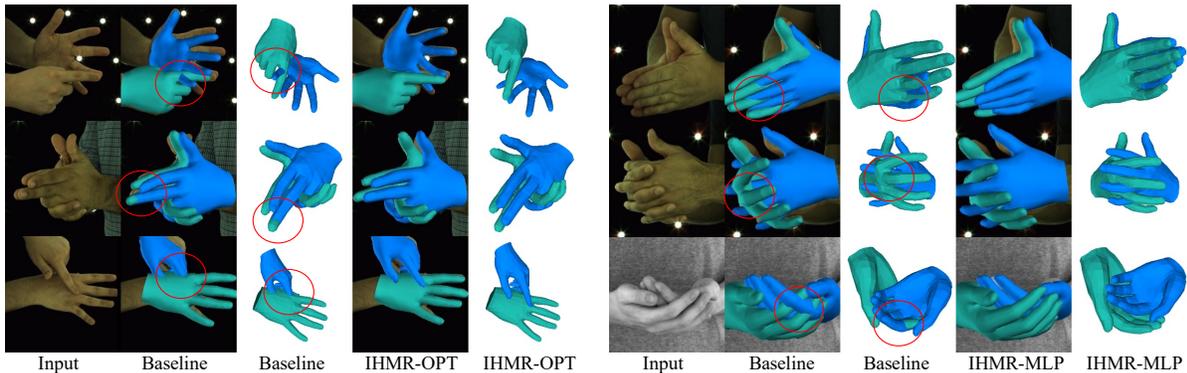


Figure 4: **Qualitative results on InterHand2.6M [22]**. Qualitative comparison between baseline model adopted from Bouk *et al.* [3] and IHMR. Collided regions are marked with red circles.

Table 2: Comparison with RGB2Hands [45] on their test set.

Method	RGB2Hands [45]	IHMR-MLP	IHMR-OPT
I-MPJPE ↓	20.02	23.41	21.30

the stronger baseline. Note that our method is advantageous than InterNet in that our reconstructed 3D meshes allow for more elaborate interactions such as "touching" and "grabbing" compared to 3D joints produced by InterNet.

Qualitative results are shown in Fig. 4. It reveals that IHMR, no matter implemented using the MLP or the optimization, can effectively diminish the collisions caused by different factors, *e.g.* relative hand positions (the first row on the left half) or finger poses (the second row on the right half). See the supplemental for more qualitative results, including the comparison with single-hand methods and the comparison between IHMR-MLP and IHMR-OPT.

Comparisons on Other Datasets. We quantitatively compare our methods with RGB2Hands [45] on the RGB2Hands evaluation set. The results are listed in Tab 2. RGB2Hands perform slightly better than our methods. We believe it is due to that RGB2Hands leverage temporal information while we only use single RGB images as inputs. For Tzionas dataset [42], we only show qualitative results in Fig. 5 due to the lack of 3D annotations. We follow the practice of InterNet [22] to train the model with the mix-



Figure 5: **Qualitative results on Tzionas dataset [42]**.

ture of InterHand2.6M data and Tzionas data. These results demonstrate the IHMR’s generalization ability.

4.3. Ablation Study

Models in this subsection are based on original IHMR described in Sec. 3 and evaluated on IH26M-Close-Inter, using I-MPJPE and AVE-P as the metrics. More ablation studies are included in the supplemental.

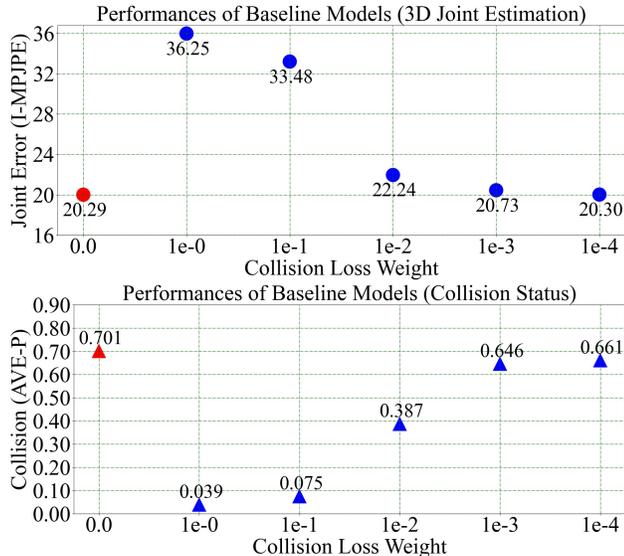


Figure 6: **Collision loss in training baseline CNN model.** We apply collision loss L_{col} in training the baseline CNN model. Red markers represent the original model trained without collision loss.

Table 3: **Design of factorized refinement.** We examine three variants of the IHMR: without factorization and without verification, and optimization from mean hand poses

Models ↓	I-MPJPE ↓		AVE-P ↓	
	OPT	MLP	OPT	MLP
Full Model	16.94	18.18	0.20	0.52
No Factorization	17.20	18.98	0.22	0.62
No Verification	16.19	26.64	0.25	0.22
OPT Mean Pose	34.74	26.22	0.03	3.67

Adopting Collision Loss in Stage-I. We conduct experiments in which the collision loss L_{col} in Eq. (7) is directly applied in training the baseline CNN model. We adjust the loss weights of L_{col} to range from 0.0 to $1e-4$, where 0.0 corresponds to the original baseline model. All models in this subsection are finetuned from a model trained on Inter-Hand2.6M without collision losses adopted. The results are depicted in Fig. 6. It is observed that although using collision loss in training baseline model can remove most of the collisions, it corrupts the 3D pose estimation, with the joint estimation errors increased by 79%. Furthermore, simply adjusting the loss weight of L_{col} cannot lead to better results with 3D hand pose estimation and collision status both improved. While decreasing loss weights of L_{col} can restrain the performance drop in the joint estimation, the collision issue becomes severe again. The results justify our design of not using collision-aware loss in the first stage but introducing it in the second stage using factorized refinement.

Design of Factorized Refinement. We evaluate three variants of IHMR. The first one is the framework without factorization, namely all parameters are refined together. The second one is performing factorized optimization without verification. The third one optimizes from mean hand

Table 4: **Influence of refining different parameters.** We evaluate shape parameters β , global hand orientation ϕ , finger poses θ and right-to-left-hand relative translation τ . Full factorized refinement with verification is applied in these experiments.

Shape	Orientation	Finger	Translation	I-MPJPE ↓		AVE-P ↓	
				OPT	MLP	OPT	MLP
✗	✗	✗	✗	20.29	20.29	0.70	0.70
✓				20.14	19.43	0.60	0.72
	✓			19.18	19.29	0.41	0.64
		✓		19.28	19.42	0.40	0.71
			✓	18.67	19.32	0.51	0.67
✓	✓			19.16	19.43	0.51	0.65
✓	✓	✓		18.38	19.42	0.38	0.63
✓	✓	✓	✓	16.94	18.18	0.20	0.52

poses, following the classical optimization methods such as SMPLify-X [28]. The results are listed in Tab. 3. Without verification, both methods fail - the 3D joint errors of IHMR-MLP increase by 47% and the average penetration of IHMR-OPT increases by 25% percent. Without factorization, the average penetration of IHMR-MLP increases by 19%. The gain is less apparent on IHMR-OPT due to its unique per-sample optimization capability, which allows IHMR-OPT to achieve optimal results more easily. Lastly, optimization starting from mean poses produces results with 3D joint errors 41% larger than IHMR-OPT does, which justifies the design of our two-stage framework.

Influence of Refining Different Parameters. Studies on the influence of refining different parameter combinations are listed in Tab. 4. Solely optimizing shape parameters provides the least performance gain. Updating finger poses can reduce the most 3D pose estimation errors, while refining orientation can diminish the most collisions. Besides, the performances can be further improved by refining more than one parameter in the second stage. Starting from the shape parameters, progressively incorporating orientation, finger poses, and translation in the whole process can consistently improve the overall performances.

5. Conclusion

We have presented the first monocular single-image based method for reconstructing 3D interacting hands with both accurate finger poses and moderate inter-collisions. To achieve this, we have proposed a novel factorized refinement that can be implemented in either traditional optimization or a feed-forward MLP. Extensive experiments on large-scale datasets demonstrate the effectiveness of our approach, which can consistently generate better 3D hand predictions despite the strong baseline such as HybriK.

Acknowledgement. This study is supported under the RIE2020 Industry Alignment Fund Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contribution from the industry partner(s). It is also partially supported by the NTU NAP grant.

Appendix

A. 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.

A.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 [22], 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 [22] 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.

A.2. Baseline CNN Model

The input images are cropped, padded, and resized to 224×224 . The encoder is a ResNet-50 [10]. 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 ($\pi \in \mathbb{R}^3$), shape parameters ($\beta = (\beta_l, \beta_r) \in \mathbb{R}^{20}$), hand orientation ($\phi = (\phi_l, \phi_r) \in \mathbb{R}^6$) finger pose parameters ($\theta = (\theta_l, \theta_r) \in \mathbb{R}^{90}$), and right-to-left-hand relative translation ($\tau \in \mathbb{R}^3$). The joint head follows the design as InterNet [22]. The whole framework, including the encoder, the parameter prediction head, and the joint head is trained end-to-end using Adam optimizer [14] with a learning rate $1e-3$. The whole model converges after 20 epochs.

A.3. Optimization-based Implementation

The optimization-based method uses Adam optimizer [14] to directly update the estimated parameters $\hat{\rho} \in \{\hat{\tau}, \hat{\phi}, \hat{\theta}, \hat{\beta}\}$. 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. A.1. We set smaller objective weights and step sizes in optimizing right-to-left-hand relative translation $\hat{\tau}$. Otherwise, the optimization tends to unreasonably increase the overall distances between collided hands and thus generate worse results.

Table A.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 $\hat{\rho} \downarrow$	γ	\mathcal{F}_{col}	\mathcal{F}_{2D}	\mathcal{F}_{3D}	\mathcal{F}_{τ}	\mathcal{F}_{reg}	\mathcal{F}_f
$\hat{\tau}$	$1e-4$	$1e-1$	$1e+1$	$1e+3$	$1e+3$	$1e-1$	0.0
$\hat{\phi}$	$1e-2$	$1e-0$	$1e+1$	$1e+3$	$1e+3$	$1e-1$	0.0
$\hat{\theta}$	$1e-2$	$1e-0$	$1e+1$	$1e+3$	$1e+3$	$1e-1$	$1e+5$
$\hat{\beta}$	$1e-2$	$1e-0$	$1e+1$	$1e+3$	$1e+3$	$1e-1$	0.0

A.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 $\hat{\tau}$, 6 for hand orientation $\hat{\phi}$, 20 for shape parameters $\hat{\beta}$ and 90 for finger poses $\hat{\theta}$. 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 [14].

A.5. HybrIK Implementation

To fill the performance gap between InterNet [22], we adopt a strong baseline following HybrIK [18]. We use the estimated 3D joints \hat{J}_{3D} from the joint head of Stage-I. To be specific, the relative right-to-left-hand translation $\hat{\tau}$ 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 $\hat{\theta}$ 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 $\hat{\beta}$ predicted from the baseline CNN model of Stage-I. We suggest to read the original paper of HybrIK [18] to have a better understanding of how HybrIK is adopted in our scenario.

B. More Quantitative Results.

Part of the training and testing data have pseudo-ground-truth MANO [29] parameters obtained from NeuralAnnot [21]. 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)

Table A.2: **Comparison with SOTA methods on InterHand2.6M using Vertex-Based Method.** We additionally evaluate on subsets of InterHand2.6M [22] 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 <i>et al.</i> [3]	21.96 / 18.88	22.55 / 18.86	-	-	21.20 / 18.81	-	-
Pose2Mesh [5]	21.76 / 18.61	22.73 / 18.91	-	-	20.82 / 18.37	-	-
BiHand [47]	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 A.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}_τ .

\mathcal{F}_{2D}	\mathcal{F}_{3D}	\mathcal{F}_τ	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 A.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 $\hat{\mathbf{J}}_{3D}$. For still joints, we still use the original pseudo-ground-truth 2D joints, $\hat{\mathbf{J}}_{2D}$.

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

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. A.2. The conclusion we draw from Tab. A.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%.

C. 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.

C.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}_τ . The results are listed in Tab. A.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.

C.2. Influence of Joint Quality.

To evaluate the convergence of the proposed factorized refinement, we conduct experiments in which pseudo-ground-truth 3D joints $\hat{\mathbf{J}}_{3D}$ 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 $\hat{\mathbf{J}}_{2D}$ 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 $\hat{\mathbf{J}}_{3D}$. (3) The effectiveness of collision removal has less correlation with the quality of $\hat{\mathbf{J}}_{3D}$.

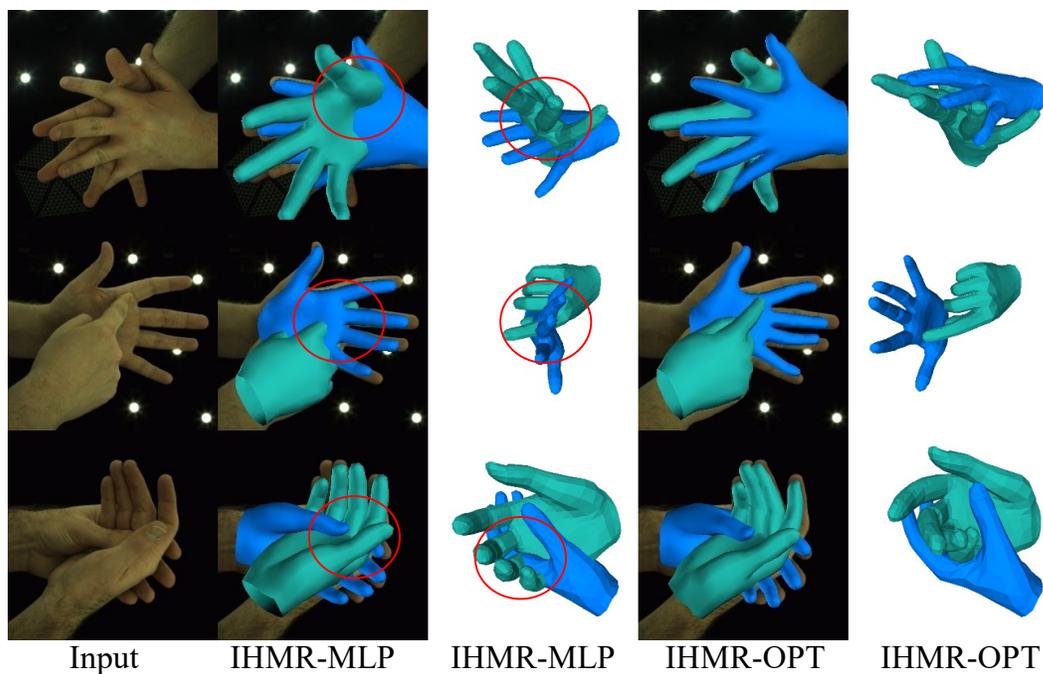


Figure C.1: **Qualitative Comparison between IHMR-MLP and IHMR-OPT.** We qualitatively compare between IHMR-MLP and IHMR-OPT. Collided regions are marked with red circles.

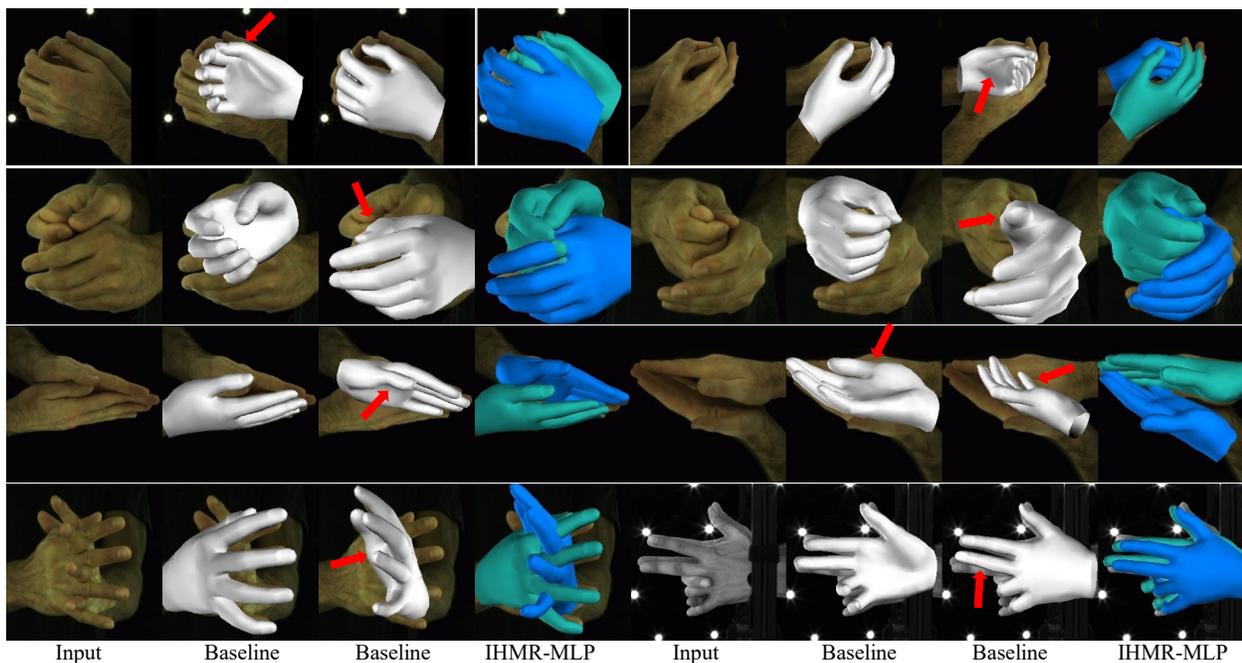


Figure C.2: **Compare with SOTA single-hand method.** In this figure, we qualitatively compare with single-hand baseline, *i.e.* Bouk [3]. From left to right are input images, predicted right hands from the baseline, predicted left hands from the baseline and predicted interacting hands from IHMR-MLP. Imprecise hand poses or finger poses generated by the single-hand method are marked with red arrows.

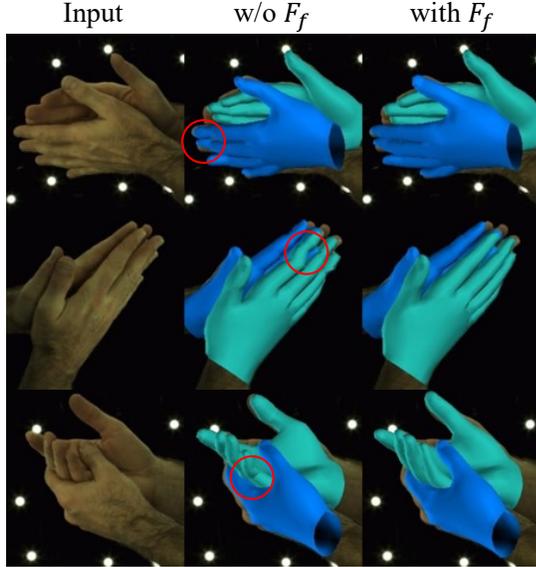


Figure C.3: **Role of \mathcal{F}_f .** In this figure, we demonstrate the role of \mathcal{F}_f plays in optimization. From left to right are input images, optimized results without using \mathcal{F}_f , optimized results with using \mathcal{F}_f . Twisted fingers generated by optimization without using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

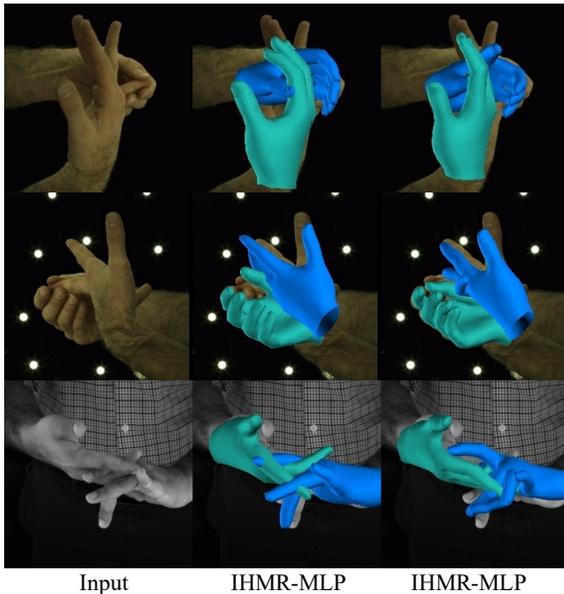


Figure C.4: **Typical Failure Cases.** Typical failure cases of IHMR are caused by challenging finger poses and collisions.

D. 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 quantitative results demonstrating the effectiveness of finger regularization \mathcal{F}_f , and typical

failure cases.

D.1. Comparison between Optimization and MLP

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

D.2. Compare with Single-hand Methods

In this subsection, we qualitatively compare IHMR with the single-hand baseline, *i.e.* Bouk *et al.* [3]. The results are shown in Fig. C.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.

D.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. C.3. It is revealed that without applying \mathcal{F}_f , optimization methods tend to generate twisted fingers as marked by red circles in Fig. C.3.

D.4. Failure Cases.

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

References

- [1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [i](#), [ii](#)
- [2] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision*, 2012. [i](#), [ii](#), [iii](#)
- [3] 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. [i](#), [ii](#), [iii](#), [vi](#), [vii](#), [x](#), [xi](#), [xii](#)
- [4] Yujun Cai, Lihao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *European Conference on Computer Vision*, 2018. [ii](#)
- [5] 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](#), [iv](#), [vi](#), [vii](#), [x](#)
- [6] Lihao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. [ii](#)
- [7] Lihao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [i](#), [ii](#)
- [8] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics*, 2020. [i](#)
- [9] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *IEEE International Conference on Computer Vision*, 2019. [i](#), [iii](#), [iv](#)
- [10] 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. [ix](#)
- [11] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. [i](#), [ii](#), [iii](#), [iv](#)
- [12] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [i](#)
- [13] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [ii](#), [iv](#)
- [14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representation*, 2015. [ix](#)
- [15] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *IEEE International Conference on Computer Vision*, 2019. [ii](#)
- [16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [ii](#)
- [17] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2020. [i](#), [ii](#)
- [18] 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. [v](#), [vi](#), [ix](#)
- [19] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. [ii](#)
- [20] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [ii](#)
- [21] 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. [vi](#), [ix](#)
- [22] 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. [ii](#), [vi](#), [vii](#), [ix](#), [x](#)
- [23] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Generated hands for real-time 3d hand tracking from monocular rgb. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. [ii](#)
- [24] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics*, 2019. [i](#), [ii](#), [iii](#)
- [25] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. [i](#)
- [26] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *British Machine Vision Conference*, 2011. [ii](#)

- [27] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012. [i](#), [ii](#), [iii](#)
- [28] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [viii](#)
- [29] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 2017. [i](#), [ii](#), [iii](#), [vi](#), [ix](#)
- [30] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3D human recovery in the wild. In *IEEE International Conference on Computer Vision*, 2019. [ii](#)
- [31] Yu Rong, Ziwei Liu, and Chen Change Loy. Chasing the tail in monocular 3d human reconstruction with prototype memory. *arXiv preprint arXiv:2012.14739*, 2020. [ii](#)
- [32] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshop*, 2021. [ii](#)
- [33] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics*, 2020. [ii](#), [iii](#)
- [34] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision*, 2020. [ii](#)
- [35] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2015. [ii](#)
- [36] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *IEEE International Conference on Computer Vision*, 2013. [ii](#)
- [37] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision*, 2018. [v](#)
- [38] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Transactions on Graphics*, 2017. [i](#), [ii](#)
- [39] Matthias Teschner, Stefan Kimmerle, Bruno Heidelberger, Gabriel Zachmann, Laks Raghupathi, Arnulph Fuhrmann, M-P Cani, François Faure, Nadia Magnenat-Thalmann, Wolfgang Strasser, et al. Collision detection for deformable objects. *Computer Graphics Forum*, 2005. [iii](#)
- [40] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM Transactions on Graphics*, 35(6):1–11, 2016. [ii](#)
- [41] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 2014. [ii](#)
- [42] 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*, 2016. [i](#), [ii](#), [iii](#), [vi](#), [vii](#)
- [43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. [ii](#)
- [44] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [ii](#)
- [45] 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*, 2020. [i](#), [ii](#), [iii](#), [vi](#), [vii](#)
- [46] Jingbo Wang, Sijie Yan, Bo Dai, and Dahua Lin. Scene-aware generative network for human motion synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2021. [i](#)
- [47] 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. [i](#), [ii](#), [vi](#), [vii](#), [x](#)
- [48] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. [ii](#)
- [49] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *European Conference on Computer Vision*, pages 346–361. Springer, 2016. [ii](#)
- [50] Jason Y Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *European Conference on Computer Vision*, pages 34–51. Springer, 2020. [i](#)
- [51] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *IEEE International Conference on Computer Vision*, 2019. [i](#), [ii](#), [v](#)
- [52] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *IEEE International Conference on Computer Vision*, 2017. [ii](#)