Decoupled Iterative Refinement Framework for Interacting Hands Reconstruction from a Single RGB Image

Pengfei Ren[1,2] Chao Wen[2] Xiaozheng Zheng[1,2] Zhou Xue[2]
Haifeng Sun[1] Qi Qi[1] Jingyu Wang[1*] Jianxin Liao[1*]
[1]State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2] PICO IDL, ByteDance, Beijing
rpf@bupt.edu.cn; {wenchao.w, zhengxiaozheng}@bytedance.com; xuezhou08@gmail.com;
{hfsun, qiqi8266, wangjingyu, liaojx}@bupt.edu.cn

Abstract

Reconstructing interacting hands from a single RGB image is a very challenging task. On the one hand, severe mutual occlusion and similar local appearance between two hands confuse the extraction of visual features, resulting in the misalignment of estimated hand meshes and the image. On the other hand, there are complex spatial relationship between interacting hands, which significantly increases the solution space of hand poses and increases the difficulty of network learning. In this paper, we propose a decoupled iterative refinement framework to achieve pixel-alignment hand reconstruction while efficiently modeling the spatial relationship between hands. Specifically, we define two feature spaces with different characteristics, namely 2D visual feature space and 3D joint feature space. First, we obtain joint-wise features from the visual feature map and utilize a graph convolution network and a transformer to perform intra- and inter-hand information interaction in the 3D joint feature space, respectively. Then, we project the joint features with global information back into the 2D visual feature space in an obfuscation-free manner and utilize the 2D convolution for pixel-wise enhancement. By performing multiple alternate enhancements in the two feature spaces, our method can achieve an accurate and robust reconstruction of interacting hands. Our method outperforms all existing two-hand reconstruction methods by a large margin on the InterHand2.6M dataset.

1. Introduction

3D hand reconstruction plays an important role in many applications, such as virtual reality (VR), augmented reality (AR), robotics, etc. With the emergence of large-scale datasets and deep learning, single-hand pose estimation and reconstruction [50, 18, 53, 34, 52, 40, 63, 6, 36, 19, 5, 59, 64, 1, 22, 46, 2, 33, 32, 9, 7, 8] have made significant progress in the past few years. Furthermore, since two hands can express richer semantics and implement more complex operations, interacting two-hand reconstruction has received a lot of attention recently. Previous work usually relies on depth cameras [3, 38, 23, 49, 37] or multi-camera systems [45], which greatly limits the application scenarios of these methods. In this paper, we focus on reconstructing interacting hands from the widely available RGB image.

Compared to the single-hand reconstruction task, reconstructing the interacting hands from a single RGB image is more challenging and far from being solved. On the one hand, severe mutual occlusion between interacting hands results in a large amount of hand area being unobservable. At the same time, the self-similar appearance brings severe am-

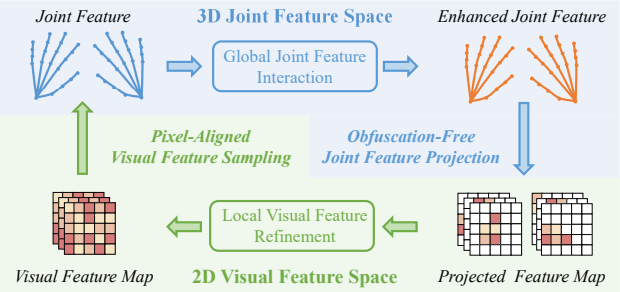

Figure 1. Decoupled Iterative Refinement. Our method extracts the visual feature map from the input RGB image, then iteratively performs visual feature refinement and joint feature interaction in a decoupled manner, and finally uses the enhanced joint feature for interacting hands reconstruction.

*Corresponding authors

biguity and confusion to the extraction of visual representations. Thus, self-occlusion and self-similarity tend to cause misalignment between the estimated hand meshes and the input image. On the other hand, the interacting hands have complex spatial relationships, and the dramatic increase in the degrees of freedom of the pose solution brings difficulties to the optimization of the network.

Some methods attempt to alleviate the interference of self-occlusion and similar appearance by utilizing some visual cues such as heatmaps [57], joint-visibility [20], and part-segmentation [13]. Nonetheless, these methods ignore the tight dependency between interacting hands. In order to better capture the spatial relationship between hands, Hampali et al. [16] extract redundant node features from the visual feature map and use a transformer to perform message passing between nodes of two hands. However, their method requires an additional association process to determine the relationship between the redundant node and the joints, which increases the difficulty of network optimization. Li et al. [24] propose to perform dense interaction between the hand vertices, so as to model the spatial relationship between the hands. At the same time, they perform dense interaction between the hand vertices and pixel features, so as to achieve image-mesh alignment. However, attention-based dense interactions are computationally expensive and suffer from the risk of overfitting.

In this paper, we decouple the difficult two-handed reconstruction task into a spatial relationship modeling problem and a pixel-level alignment problem, which can be handled in a simple but efficient way in specialized spaces, respectively. As shown in Fig 1, our method explicitly defines two feature spaces, 2D visual feature space and 3D joint feature space. In 3D joint feature space, we represent hand information by compact joint features. We adopt a Graph Convolution Network (GCN) [60] and a transformer [61] to model the intra- and inter-relationships of two hands, respectively. In 2D visual feature space, we adopt the visual feature map to represent two hands information and enhance local visual features by fusing joint features in an obfuscation-free way. We communicate the two spaces through an unambiguous 2D-3D coordinate projection relationship. Performing long-range relational modeling in joint feature space is computationally friendly. It can take advantage of skeletal structure information, which reduces the difficulty of network optimization. At the same time, joint features with global information can provide strong disambiguation clues for local visual features, alleviating the information loss caused by self-occlusion and the ambiguity caused by self-similarity.

Experiments show that our method significantly outperforms State-Of-The-Art (SOTA) methods on the InterHand2.6M dataset [24]. At the same time, we also show qualitative images and videos (see supplementary material) of our method on multiple in-the-wild samples [51, 54, 4, 44], from which we can observe that our method has a strong generalization ability. Code is available at: https://github.com/PengfeiRen96/DIR.

Our contributions can be summarized as follows:

- We propose a decoupled iterative refinement framework for reconstructing interacting hands. Our method achieves pixel-level mesh-image alignment while efficiently modeling the spatial relationship of the hands.
- We model the spatial relationship of two hands through compact and semantically explicit joint nodes, which is computationally friendly and can utilize the priors of hand bone structure.
- We propose an obfuscation-free way to project joint features into visual feature space, which alleviates the ambiguity caused by self-similarity and the absence of visual cues caused by self-occlusion.
- Our method outperforms recent SOTA methods by a large margin and shows a strong generalization ability for the in-the-wild images.

2. Related Works

2.1. Single Hand Reconstruction

Early single-hand reconstruction work relied on depth data [18, 53, 34, 50, 52, 40], but with the emergence of large-scale datasets and the development of deep learning, RGB-based hand reconstruction has made great progress. Pioneering RGB-based work focus on only estimating hand pose from input [63, 47, 19, 6]. With the proposal of parametric hand models, some work [62, 59, 5, 31, 21] attempt to reconstruct hands directly from RGB images using MANO models. However, it is challenging to predict the parameters of the hand model from a single RGB image, which leads to difficulties in network optimization and is prone to mesh-image misalignment [48]. To alleviate these problems, some works propose to use GCN [14, 22, 11] or transformer [26, 25, 10] to directly estimate the coordinates of the vertices of the hand mesh. However, reconstructing the hand without prior knowledge can easily lead to the collapse of the predicted mesh, even if these methods adopt some constraint terms to keep the generated mesh smooth. Therefore, in order to keep the estimated hand model reasonable, we adopt a parametric hand model and alleviate the mesh-image misalignment through the iterative 3D spatial relationship modeling and the 2D feature enhancement.

2.2. Interacting Hand Reconstruction

Interacting hand reconstruction is a very challenging problem. Some pioneering work [3, 38, 23] fit a parametric hand model with observed depth data by optimizing an

energy function. These methods tend to be trapped in local optima and are computationally expensive. A common approach is to train a deep neural network to predict some visual cues, such as segmentation[49, 37], pixel-mesh correspondence map [37, 54] or the dense relative depth map of the interacting hands [54], to reduce the search space of hand poses and the optimization difficulty of the energy function. However, this hybrid approach cannot be trained in an end-to-end manner and the optimization process may still fall into local minima. Recently, Smith et al. [45] propose a multi-view camera system, which can reconstruct high-fidelity interacting hand motions. However, this method requires custom-built dedicated hardware and the algorithm is time-consuming.

In recent years, with the proposal of the large-scale interacting hand dataset InterHand2.6M [35], great progress has been made in single RGB-based 3D interacting hands reconstruction. Moon et al. [35] extend the single-hand pose estimation method to the two-hand interacting scenario, which predicts the 2.5D heatmap of the two hands simultaneously. Some works improve the accuracy of the interacting hand estimation by incorporating some visual cues such as joint-visibility [20] and part-segmentation [13]. Rong et al. [43] propose a two-stage framework to alleviate the collision problem between the hands. In order to reduce the mutual interference between interacting hands, Zhang et al. [57] propose to use heatmaps to make the network focus on specific regions, and Meng et al. [30] adopt an erase mechanism to convert the two-hand image into two single-hand images. However, these methods do not adequately model the dependencies between the two hands. Hampali et al. [16] adopt the transformer to model the interaction between two hands, which is robust but still hard to mitigate misalignment between the estimated hand pose and the image. Li et al. [24] progressively enhance mesh vertex features with image features while performing information interaction between two-hand meshes, which is helpful for mesh-image alignment. However, performing dense mesh-mesh and mesh-images interactions is computationally complex and prone to overfitting.

2.3. Pixel-level Alignment

Estimating pixel-aligned 3D mesh directly from a single RGB image is quite challenging, either by estimating parametric models or by directly predicting mesh vertex coordinates. Wang et al. [55] and Wen et al. [56] propose to use the camera intrinsic matrix to obtain the perceptual features of each 3D mesh vertex from the 2D image features according to the 3D-2D coordinate relationship. Further, Zhang et al. [58] and Tang et al. [48] extend this strategy to extract human body mesh and hand mesh features from visual features, respectively, in order to obtain a more accurate mesh-image alignment. In addition to sampling mesh features using coordinate relations, another way [26, 24] is to use a transformer to perform densely interaction between image features and vertex features. However, this method is computationally expensive and relatively sensitive to self-occlusion (the correctness of the interaction relationship is affected by the quality of the feature itself). In particular, the above two methods only perform feature enhancement in the 3D vertex space, ignoring the role of the 2D feature space. Specifically, the local receptive field mechanism of 2D convolution operation provides intrinsic inductive bias, which can efficiently and effectively utilize local features for pixel-level refinement. Our method projects joint features with global information back into the 2D visual feature space, which further provides strong cues for 2D convolution to achieve more accurate pixel-level alignment.

3. Method

In this paper, we propose a decoupled iterative refinement module for interacting two-hand reconstruction from a single RGB image. As shown in Fig. 2, we adopt an encoding-decoding network structure. The encoder extracts multi-scale visual features from the input image, and uses global features to estimate the initial hand meshes and the relative offset of two hands. Then, the decoder progressively enhance the visual feature maps and refine the hand meshes and the relative offset. During the decoding process, we iteratively perform two-hands spatial relationship modeling and visual feature refinement in a decoupled manner.

3.1. Encoder and Initial Estimation

We adopt a ResNet-50 [17] pretrained on ImageNet [12] as the encoder, from which we can obtain multi-scale visual features $\{\mathbf{F}_n \in \mathbb{R}^{C_n \times H_n \times W_n}\}_{n=0}^{n=N-1}$, where N, C_n, H_n, W_n represent the number of visual feature scales, the channel dimension, height and width of the n-th feature map respectively. In general, direct regression of vertex coordinates can achieve higher accuracy mesh prediction [48, 26, 14, 22]. However, this method is prone to generate collapsed and unreasonable hand mesh, so the robustness is relatively poor. Therefore, we use the global features extracted by the encoder to regress the parameters of a parameterized hand model MANO [42] directly and then improve the accuracy of the initial meshes through subsequent iterative refinement. In particular, the two hands should have their own unique features, so we adopt a simple and lightweight attention module to separate the features of the two hands from the highest-level image feature map $\mathbf{F}_{N-1}$. Taking the left hand as an example, the left hand global feature $\mathbf{G}_{left} \in \mathbb{R}^{C_{N-1}}$ can be obtained by an attention map $\mathbf{A}_{left} \in \mathbb{R}^{1 \times H_{N-1} \times W_{N-1}}$ as follow:

$$\mathbf{A}_{left} = Sigmoid\left(Conv_{left}\left(\mathbf{F}_{N-1}\right)\right), \tag{1}$$

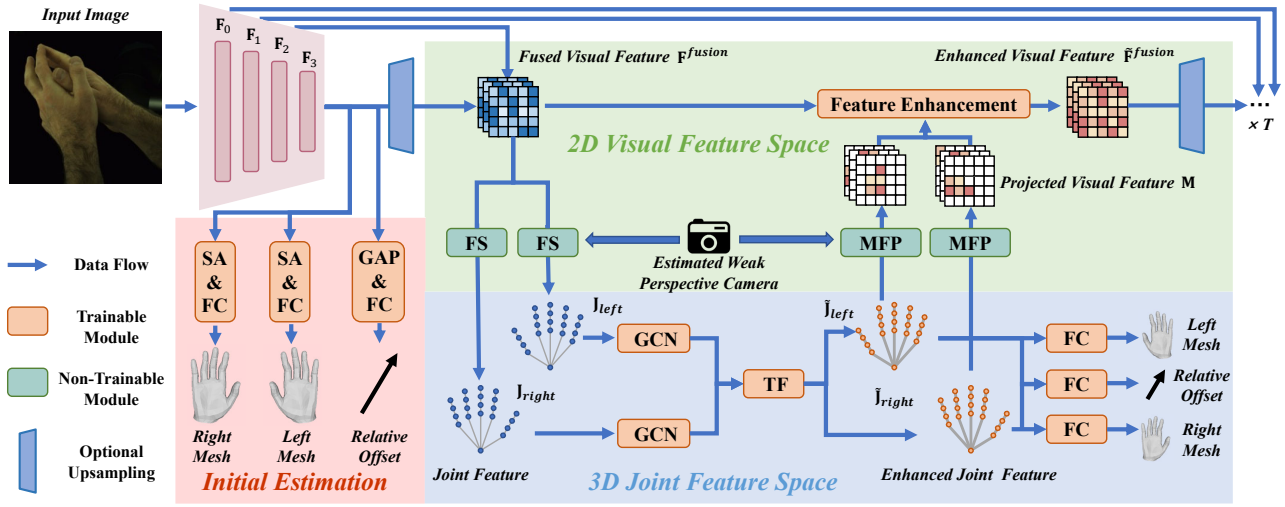

Figure 2. Our framework. We utilize the global features extracted by the encoder to predict the initial hand meshes and the relative offset of the hands. Then, the decoder gradually fuses the multi-scale visual feature maps from the encoder and refines the hand meshes and the relative offset. 'SA' and 'GAP' represent spatial attention and global average pooling, respectively. 'FS' and 'MFP' represent joint-wise feature sampling and multi-plane feature projecting, respectively. 'TF' represents a multi-layer transformer. In particular, since our method limits the resolution of feature maps to a maximum of 32×32 during decoding, we only need to adopt the upsampling in the first refinement stage. In addition, for feature maps from the encoder with a resolution greater than 32×32, we change its resolution by downsampling.

$$\mathbf{G}_{left} = AvgPool\left(\mathbf{A}_{left} * \mathbf{F}_{N-1}\right). \tag{2}$$

Then, we estimate model parameters ($\theta_{left} \in \mathbb{R}^{62}$ and $\theta_{right} \in \mathbb{R}^{62}$) and weak perspective camera parameters ($\mathbf{P}_{left} \in \mathbb{R}^{3}$ and $\mathbf{P}_{right} \in \mathbb{R}^{3}$) of the two hands from $\mathbf{G}_{left}$ and $\mathbf{G}_{right}$ through Fully Connected (FC) layers, respectively. At the same time, we predict the relative offset $\mathbf{O} \in \mathbb{R}^{3}$ of the two hands from the features $\mathbf{F}_{N-1}$ through a 2D average pooling and a FC layer.

3.2. Decoder and Iterative Refinement

Modeling the 3D spatial relationship of the two hands and aligning the estimated mesh with the observed 2D image are two major challenges for interacting hands reconstruction. We address these two problems in the 2D image feature space and 3D joint feature space, respectively, in a decoupled manner. In a single refinement stage, first, we extract joint features from the 2D image feature map according to the 3D-to-2D coordinate relationship. Then, in the 3D joint feature space, we perform intra- and inter-hand information interaction to capture the complex spatial dependencies between two hand joints. Finally, we project the joint-wise features with global context information back to the 2D image space in an unobfuscated way, which provides strong disambiguation clues for local visual features refinement. In particular, we take a total of T refinement stages.

3.2.1 Constructing Joint Feature

For the t-th refinement stage, given the feature map from the previous decoding layer and the skipped image feature map $\mathbf{F}_{N-t-2}$ from the corresponding encoder layer, we concatenate them together and obtain the fused feature map $\mathbf{F}^{fusion}$ by a 1×1 convolution layer. Then, similar to [48, 58, 55, 39, 31], we obtain joint-wise visual feature $\mathbf{J}^{visual} \in \mathbb{R}^{C\times J}$ from $\mathbf{F}^{fusion}$ via a 3D-to-2D coordinate projection and a bilinear interpolation around each projected joint coordinate, where J, C represent the joint number and channel dimension of the joint feature. The 3D hand joint coordinates are obtained from the 3D hand mesh predicted in the previous stage. In particular, the projection here is determined by the estimated weak perspective camera, so our method does not require camera intrinsics. In addition, we encode the estimated joint coordinates into coordinate features $\mathbf{J}^{coord} \in \mathbb{R}^{C\times J}$ through a FC layer. With the joint-wise visual feature and coordinate features, we can obtain the initial joint features $\mathbf{J} = \mathbf{J}^{coord} + \mathbf{J}^{visual}$. We perform joint feature extraction for two hands independently. In particular, we use the predicted relative offset $\mathbf{O}$ to move the left-hand and right-hand coordinates in the same 3D space.

3.2.2 Modeling Spatial Relationship in 3D Space

For the 3D spatial relationships, our method mainly focuses on two parts, one is the joint dependencies of a single hand, and the other is the spatial context relationships between

two hands. First, there are explicit dependencies between the joints of a single hand. Utilizing the intrinsic dependency of the hand structure to perform information interaction between joints can reduce the difficulty of network optimization and alleviate the interference of low-quality features. For example, when a joint lacks explicit visual cues due to occlusion, we can infer its location based on its related joints such as its parent and child joints. Therefore, we utilize a GCN [60] to perform intra-information interaction between the joint nodes of a single hand based on the skeletal structure. Meanwhile, for the tightly interacting hands, there are more complex and flexible spatial relationships between the joints of the two hands. Therefore, we adopt a multi-layer transformer [61] to model the relationship between two-hand joints. By using the GCN and the transformer for information interaction, we can obtain enhanced joint features $\widetilde{\mathbf{J}} \in \mathbb{R}^{C\times J}$ with global information, which are used for MANO parameter prediction and subsequent visual feature enhancement.

Joint nodes have clear semantics, which can take advantage of hand bone structure during interaction and reduces the optimization difficulty of the network. Compared with using redundant nodes to model the two-hand relationship [16], our method can avoid the extra node-to-joint assignment. Compared with performing information interaction between the mesh vertices of two hands [24], our method is computationally efficient and can avoid overfitting.

3.2.3 Enhancing Visual Feature in 2D Space

Benefiting from the local receptive field mechanism and intrinsic inductive bias, 2D convolution can capture local structures efficiently and effectively. However, convolution operations are difficult to model long-range relationships, even when stacking multiple convolutional layers [29]. Therefore, 2D visual features are susceptible to self-occlusion or self-similar appearance interference. To alleviate this problem, we project joint features with global information back to 2D image features, which can provide strong disambiguation clues to local visual features. Furthermore, in order to prevent feature confusion when different joint features are projected back to the same or near pixel positions, we propose a Multi-plane Feature Projecting (MFP) mechanism. Specifically, we independently project each joint feature to a feature map $\mathbf{M}_j \in \mathbb{R}^{C\times H\times W}$ and then concatenate $2J$ feature maps as the final projected feature map $\mathbf{M} \in \mathbb{R}^{(2J\times C)\times H\times W}$. Previous work [24, 48, 58, 55] focusing on mesh-image alignment only extracts vertex features unidirectionally from 2D image features and performs refinement in 3D mesh space. These methods do not enhance visual features and ignore the role of pixel-level refinement.

3.3. Loss Functions

The loss function consists of three parts, including a MANO loss, an offset loss, and a pixel-wise loss.

MANO Loss. We supervise the hand joints and meshes predicted by the network. Similar to previous methods [24], we supervise the root-relative 3D joint coordinates $\mathbf{C}^{3D}$, 2D joint coordinates $\mathbf{C}^{2D}$, 3D coordinates of mesh vertices $\mathbf{V}^{3D}$, and 2D coordinates of mesh vertices $\mathbf{V}^{2D}$ as follows:

$$L_{joint} = \sum_{i=0}^{T-1}\sum_{j=0}^{J-1} L1\left(\mathbf{C}_{i,j}^{3D}, \mathbf{C}_{i,j}^{3D,gt}\right) + L1\left(\mathbf{C}_{i,j}^{2D}, \mathbf{C}_{i,j}^{2D,gt}\right), \tag{3}$$

$$L_{mesh} = \sum_{i=0}^{T-1}\sum_{j=0}^{V-1} L1\left(\mathbf{V}_{i,j}^{3D}, \mathbf{V}_{i,j}^{3D,gt}\right) + L1\left(\mathbf{V}_{i,j}^{2D}, \mathbf{V}_{i,j}^{2D,gt}\right), \tag{4}$$

where $L1$ represents the smooth L1 loss [15, 41]; T, J, V represents the number of iterative refinements, the number of joints and the number of mesh vertices, respectively. Besides, similar to previous methods [24, 48, 32], we also adopt a normal consistency loss and an edge length consistency to maintain the smoothness of the estimated mesh.

Offset Loss. Accurately estimating the offset between the two hands is important for modeling the spatial relationship of the hands. Therefore, we supervise the offsets of the joints of two hands.

$$L_{offset} = \sum_{i=0}^{T-1} L1(\mathbf{O}_i, \mathbf{O}_i^{gt}) \tag{5}$$

Pixel-wise Loss. Similar to [58], we utilize auxiliary tasks for pixel-level supervision, enhancing the reliability of visual features. Specifically, we predict two-hand segmentation and dense correspondence maps [24, 37, 54], and supervise them with mean square error (MSE) loss.

4. Experiment

4.1. Implementation Details

We train and evaluate our method on a single server with an NVIDIA A100 Tensor Core GPU. The network is implemented within PyTorch. We train our network using the AdamW [28] optimizer with an initial learning rate of 3e-4 and a cosine decay learning rate schedule [27]. The whole training process takes 50 epochs with a batch size of 64. We perform data augmentation including random rotation, random scaling, random translation, random horizontal flipping and motion blur. We crop out the hand region based on the 2D coordinates of the hand vertices and resize it to 256 $\times$ 256. More details about network structure and training details are provided in the supplementary material.

ID	IR	GCN	TF	SFP	MHP	MFP	MPJPE	MPVPE	MIAA
1							12.44	12.11	7.41
2	√						11.30	11.03	6.60
3		√					10.94	10.75	6.42
4	√		√				10.98	10.80	6.45
5		√	√				10.73	10.53	6.31
6				√			10.68	10.45	6.27
7	√	ALL			√		10.65	10.44	6.27
8						√	10.49	10.26	6.18

Table 1. We report the MPJPE (mm), MPVPE (mm) and MIAA (pixel) on InterHand2.6M dataset. 'AM' represents using attention map to split features of the left and right hands. 'IR' stands for adopt iterative refinement. 'SFP' and 'MHP' represent the single-plane feature projection and the multi-plane heatmap projection.

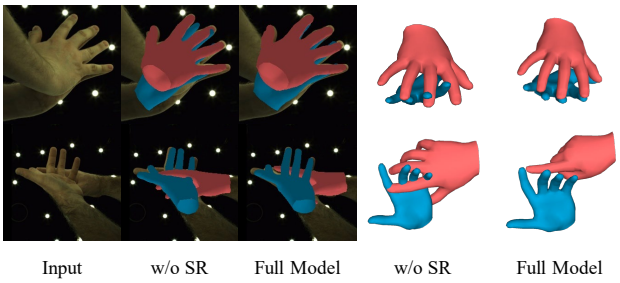

Figure 3. Qualitative ablation study. 'w/o SR' means removing the GCN and the transformer from our full model. In addition to the mesh that overlaps the input image, we also show the estimated two-hand mesh from the other viewpoint.

4.2. Datasets

InterHand2.6M. We majorly conduct experiments on InterHand2.6M [35], which provides multi-view RGB images with two-hand mesh and joint 3D annotation. Instead of using multi-view information, we treat all images as single-view images. This dataset is very challenging, it contains complex two-hand interaction poses and covers large-scale perspective changes. Following the practice of [24], we use the 5 FPS version of the released data and only use the interacting two-hand data. Specifically, it consists of 366K training images and 261K testing images.

In-the-wild Datasets. We conduct qualitative experiments on RGB2Hands dataset [54], EgoHands dataset [4], 100DOH dataset [44] and the dataset proposed by Tzionas et al. [51]. These datasets have complex interacting hand samples, diverse backgrounds, realistic lighting conditions and varying image quality, which can provide a comprehensive evaluation of the generalization ability of our approach.

4.3. Evaluation Metrics

First, we adopt Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE) to measure the accuracy of the pose and shape of the estimated hand meshes. For a fair comparison, we follow the previous work [35, 16] that use the wrist as root to perform joint alignment and scale the prediction according to the ground-truth bone length when evaluating. In particular, in qualitative experiments, we do not perform root joint alignment and scaling. second, we adopt the Mean Relative-Root Position Error (MRRPE) to evaluate the root-relative the translation between the hands. Third, to evaluate the Mesh-image Alignment Accuracy (MIAA), we calculates the 2D distance in image pixels between the projected ground truth vertices and the predicted vertices. In addition, we report the Percentage of Correct Keypoints (PCK) and Area Under the Curve (AUC) between 0 and 50 millimeters.

4.4. Ablation Study

Basic Network. In this section, we experiment with different ways of predicting initial MANO parameters. Here, our network does not adopt iterative refinements or decoders. As shown in Table 1, the basic method (ID 1) regresses the MANO parameters and the relative offset directly, which performs poorly. Similar to [58], we extract pixel-aligned features from visual features based on joint coordinates and perform two iterations of correction (ID 2), which can significantly improve the performance of the network. In subsequent experiments, we adopt two refinement stages, by default.

Modeling Spatial Relationships In this section, we evaluate the impact of modeling intra- and inter-hand relationships on performance. Compared with not modeling any spatial relationship (ID 2), either using GCN (ID 3) for information interaction between single-hand joints or using a transformer (ID 4) for spatial relationships modeling between two-hand joints can bring a significant performance improvement. Furthermore, using GCN and transformer at the same time (ID 5) can achieve the best performance.

In Fig. 3, we show the importance of Spatial Relationship (SR) modeling between hands. On the one hand, abandoning spatial interactions of the two hands leads to severe intersections and collisions, especially for invisible parts (row1). On the other hand, for the samples with visible interaction regions but complex interaction pose, abandoning two-hand information interaction also leads to the wrong spatial relationship between interacting joints (row2).

Joint Feature Projection Joint features with global information can provide strong disambiguation cues for local visual features. First, if all joint features are projected into a single plane (ID 6), the features of different joints will be confused, so the performance also drops to a certain extent compared with ID 8. Converting the joint coordinates into multiple heatmaps (ID 7) avoids the confusion of different joints, but the information passed to the visual features is limited (only position information), so the performance is

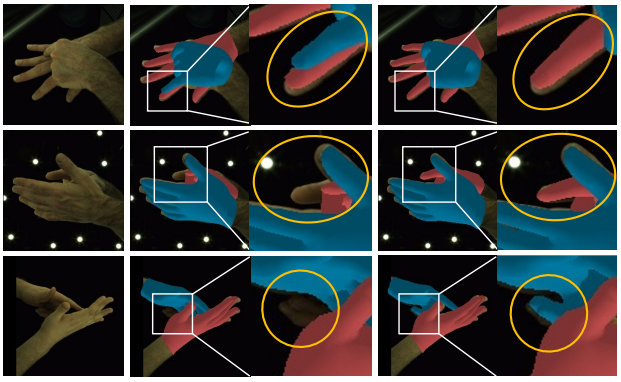

Figure 4. Qualitative Ablation study. 'w/o MFP' means removing the MFP process from our full model.

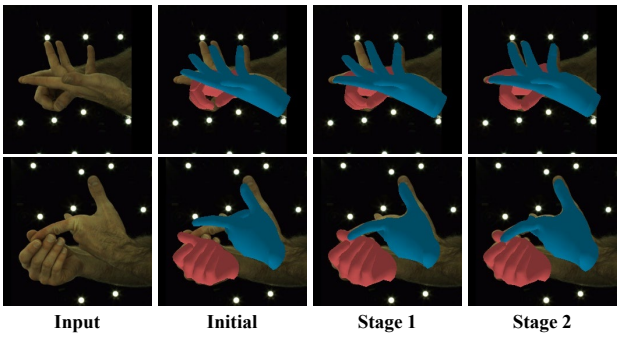

Figure 5. From left to right, we show that the initial hand meshes are gradually corrected. A more accurate spatial relationship modeling can provide stronger disambiguation cues for 2D feature refinement in 2D visual space; better 2D visual features help to construct more representative joint nodes in 3D joint space.

not good enough. Adopting MFP (ID 8) not only avoids feature confusion but also preserves the information of node features well, which achieves the best performance.

In Fig. 4, we show the effect of MFP. First, we observe that adopting MFP can alleviate the ambiguity caused by the self-similarity between hands (row1), which indicates that the projected two-hand information provides strong disambiguation cues for local visual features. Second, when there is severe slef-occlusion between hands, the projected information can enhance the local visual feature of unobservable regions, significantly improving the robustness of the estimation for the occluded hand (row2). We also observe that adopting MFP can help the network to focus on some areas that are easily ignored, such as dark regions (row3).

Iterative Refinement In this section, we explore the effect of the number of refinement stages on the performance. With one, two, and three refinement stages, MPJPE can reach 10.95 mm, 10.49 mm, and 10.44 mm, respectively, which are about 12.0%, 15.7%, and 16.1% higher than the initial estimation. A better balance of speed and accuracy can be achieved with two-stage refinement (31 FPS), while higher performance can be achieved with three-stage refinement (23 FPS). As shown in Fig 5, our method can effectively correct initial estimates with misaligned meshes or wrong spatial relationships between hands.

Method	MPJPE	MPVPE	MIAA	MRRPE
InterNet [35]	14.42	-	-	30.08
InterShape [57]	12.54	12.26	7.40	-
KPT [16]	12.42	-	-	29.17
IntagHand [24]	12.40	12.09	7.11	30.40
Ours	**10.49**	**10.26**	**6.18**	**28.98**

Table 2. We report MPJPE (mm), MPVPE (mm), MIAA (pixel) and MRRPE (mm) on InterHand2.6M.

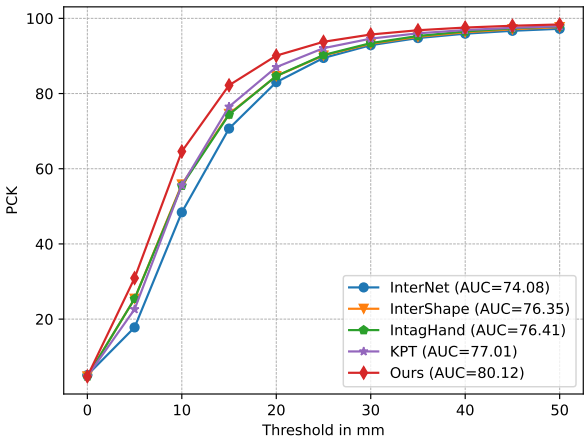

Figure 6. Comparison with SOTA methods on InterHand2.6M.

4.5. Comparisons with State-of-the-arts

We compare our method SOTA two-hand reconstruction methods, including InterNet [35], InterShape [57], IntagHand [24] and KeyPoint Transformer (KPT) [24]. For the fairness of comparison, we re-evaluate these methods in the same way, that is, using the ground-truth of wrist joint for hand alignment and scaling the estimated results according to the ratio of the predicted bone length to the ground-truth bone length. In particular, we evaluate these methods using their released source code and model weights. As shown in Table 3, our method significantly outperforms all two-hand reconstruction methods. First, compared with SOTA two-hand estimation method IntagHand [24], our method improves by 15.40% (10.49 mm vs. 12.40 mm) and 15.14% (10.26 mm vs. 12.09 mm) in MPJPE and MPVPE. Second, since our method explicitly models the two hand spatial relations in the 3D joint feature space, our method shows a significant advantage in MRRPE. Third, our method exhibits better pixel-aligned properties, achieving the lowest

Method	MPJPE (MS)	MPVPE (MS)	MPJPE (M)	MPVPE (M)
InterNet [35]	12.79	-	12.94	-
KPT [16]	9.17	-	9.61	-
InterShape [57]	8.88	9.15	10.51	10.95
IntagHand [24]	8.79	9.03	10.14	10.59
Ours	**7.51**	**7.72**	**8.70**	**9.06**

Table 3. We report MPJPE-MS (mm), MPVPE-MS (mm), MPJPE-M (mm), MPVPE-M (mm) on InterHand2.6M.

MIAA. In addition, as shown in Fig. 6, our method outperforms previous methods at almost all error thresholds and has the highest AUC.

Meanwhile, we report MPJPE and MPVPE using the middle finger MCP joint for alignment while using scaling (the evaluation method is the same as IntagHand [24]), called MPJPE-MS and MPVPE-MS, respectively. We also report MPJPE and MPVPE using MCP alignment without scaling, referred to as MPJPE-M and MPVPE-M, respectively. As shown in Table 3, evaluating with different settings can lead to significantly different performances. For example, using the middle finger MCP joint alignment can significantly improve the performance of the network. However, regardless of the evaluation strategy, our method outperforms SOTA methods by a large margin under all metrics.

In addition, we adopt Interpenetration Volume (IV) to measure the degree of collision between two hands. IV is obtained by voxelising the hand model using a voxel size of 0.5 cm. Compared with previous SOTA method IntagHand [24], without explicitly modeling the collision between the hands, our method has a lower IV (4.14 cm^3 Vs. 4.37 cm^3). This shows that our method better models the spatial relationship between hands.

4.6. Qualitative Results

We present the qualitative results of our method on the Interhand2.6M in Fig. 7. Compared with IntagHand [24], our method can avoid the collapse of the estimated mesh (row 1, row 2); without an explicitly collision constraint, our method can better avoid unreasonable intersections between hands (row 3, row 4). This shows that the joint-based information interaction can help our model capture the spatial relationship between the joints. Meanwhile, our method achieves better mesh-image alignment (row 2, row 3). In particular, when a hand is almost completely occluded (row 5), our method can also infer the pose of the occluded hand based on fine-grained visual cues and global information. We provide a comparison with IntagHand on live video in the supplementary material.

We also show some failure examples of our method on the Interhand2.6M. As shown in Fig. 8, for cases with

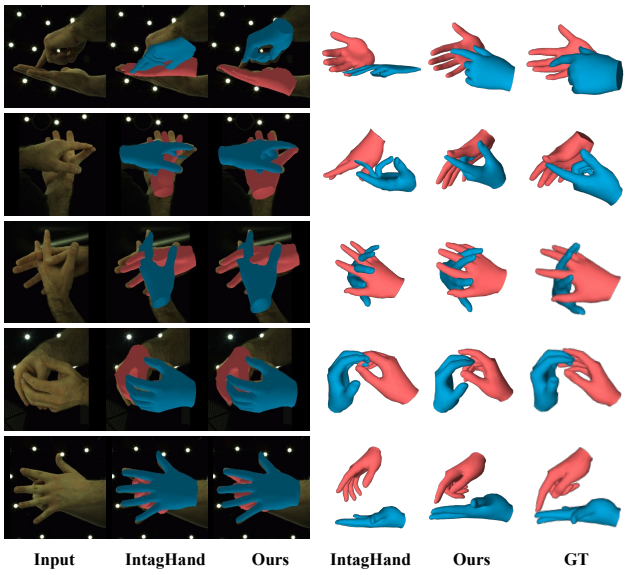

Figure 7. Qualitative results of IntagHand [24] and our method on InterHand2.6M dataset. Left: frontal view to evaluate the alignment accuracy; Right: another view to observe the plausibility of the estimated two hand meshes.

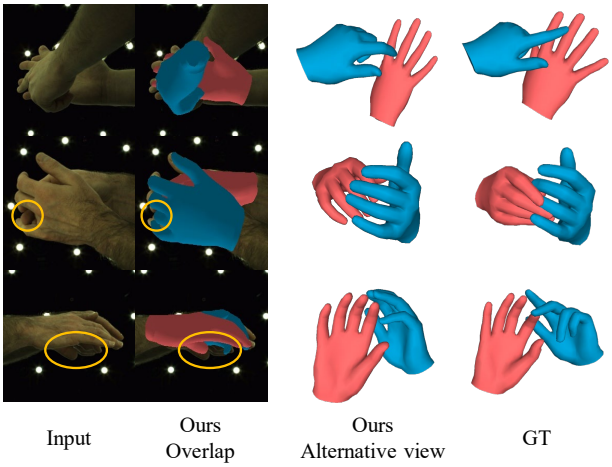

Figure 8. Failure examples on InterHand2.6M [35]. We highlight the region where the reconstruction is wrong with a yellow circle. 'GT' means the ground truth.

severe self-occlusion and fine-grained interactions (row1), our method does not achieve accurate mesh-image alignment and has a wrong understanding of the two-hand relationship. Second, for examples with severe self-occlusion, when the target joint has few observable pixels (row2) or the observable region is dark (row3), our method cannot reconstruct the corresponding region accurately. In conclusion, our method may fail when multiple conditions such as self-occlusion, tight interaction, blurring or shadowing occur simultaneously or in combination.

Similar to [24, 57], we also demonstrate the general-

Figure 9. Qualitative results on in-the-wild images. Each row from top to bottom corresponds to RGB2Hands dataset[54], the dataset proposed by Tzionas et al. [51], EgoHands dataset [4] and 100DOH dataset [44]. In each part, the left is the input image, the middle is the result of IntagHand [24], and the right is the result predicted by our method.

ization ability of our method on in-the-wild images. It is worth mentioning that we only use InterHand2.6M dataset for training without any additional pre-training on synthetic hand datasets or fine-tuning on other hand datasets. As shown in Fig. 9, our method has a strong generalization ability for different viewpoints such as third or egocentric viewpoints. At the same time, as shown in the third row and the last row, our method is more robust to different backgrounds and lighting conditions. Since our method is able to enhance visual features with global information, our method is also relatively robust to object perturbations. In particular, compared with IntagHand [24], which has strong generalization ability, our method also shows obvious advantages, especially in that our method can better maintain hand structure. We present more qualitative results in the supplementary material.

5. Conclusion

In this paper, we propose the decoupled iterative refinement framework to reconstruct interacting hands from a single RGB image. In order to efficiently model the spatial dependencies between the two hands, we adopt the GCN and the transformer to perform intra- and inter-hand information interaction in the 3D joint feature space. To achieve better alignment of the estimated mesh and observed images, we project the joint features with global information into the visual feature space in a confusion-free manner, which provides strong disambiguation cues for visual features, alleviating self-occlusion and self-similarity problems. Ablation experiments demonstrate that the decoupled iterative refinement can effectively solve two major challenges in interacting hands reconstruction, namely, modeling the complex hands spatial relationships and visual feature deconfusion. Quantitative experiments on InterHand2.6M show that our method outperforms the previous SOTA by a large margin. Meanwhile, experiments on in-the-wild images demonstrate that our method has a strong generalization ability.

Limitation and Future Work. Our method does not explicitly model collisions between hands, so even with the intra- and inter-hand relationships modeling, intersections between hands still occur, in some cases. Besides, our method does not fully utilize the estimated 3D mesh information. Mesh information may helpful for a fine-grained understanding of the relationship between hands. Finally, to achieve finer-grained mesh-image alignment, highfidelity parametric hand models may be beneficial.

Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants (62071067, 62171057, 62201072), in part by the Ministry of Education and China Mobile Joint Fund (MCM20200202), Beijing University of Posts and Telecommunications-China Mobile Research Institute Joint Innovation Center, in part by the Project funded by China Postdoctoral Science Foundation (2023TQ0039).

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 *CVPR*, pages 1067–1076, 2019. 1

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *CVPR*, pages 6121–6131, 2020. 1

[3] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, pages 640–653. Springer, 2012. 1, 2

[4] Sven Bambach, Stefan Lee, David J Crandall, and Chen Yu. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. In *ICCV*, pages 1949–1957, 2015. 2, 6, 9

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, pages 10843–10852, 2019. 1, 2

[6] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, pages 666–682, 2018. 1, 2

[7] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *ICCV*, pages 12929–12938, 2021. 1

[8] Xingyu Chen, Yufeng Liu, Yajiao Dong, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, pages 20544–20554, 2022. 1

[9] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *CVPR*, pages 13274–13283, 2021. 1

[10] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, pages 342–359. Springer, 2022. 2

[11] 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 *ECCV*, pages 769–787. Springer, 2020. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 3

[13] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *3DV*, pages 1–10. IEEE, 2021. 2, 3

[14] Liuhao 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 *CVPR*, pages 10833–10842, 2019. 2, 3

[15] Ross Girshick. Fast r-cnn. In *ICCV*, pages 1440–1448, 2015. 5

[16] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, pages 11090–11100, 2022. 2, 3, 5, 6, 7, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, June 2016. 3

[18] Weiting Huang, Pengfei Ren, Jingyu Wang, Qi Qi, and Haifeng Sun. Awr: Adaptive weighting regression for 3d hand pose estimation. In *AAAI*, volume 34, pages 11061–11068, 2020. 1, 2

[19] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 1, 2

[20] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *CVPR*, pages 11189–11198, 2021. 2, 3

[21] Deying Kong, Linguang Zhang, Liangjian Chen, Haoyu Ma, Xiangyi Yan, Shanlin Sun, Xingwei Liu, Kun Han, and Xiaohui Xie. Identity-aware hand mesh estimation and personalization from rgb images. In *ECCV*, pages 536–553. Springer, 2022. 2

[22] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 1, 2, 3

[23] Nikolaos Kyriazis and Antonis Argyros. Scalable 3d tracking of multiple interacting objects. In *CVPR*, pages 3430–3437, 2014. 1, 2

[24] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *CVPR*, pages 2761–2770, 2022. 2, 3, 5, 6, 7, 8, 9

[25] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, June 2021. 2

[26] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *ICCV*, pages 12939–12948, 2021. 2, 3

[27] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5

[28] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[29] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. Understanding the effective receptive field in deep convolutional neural networks. *NIPS*, 29, 2016. 5

[30] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *ECCV*, 2022. 3

[31] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3d hand pose estimation for whole-body 3d human mesh estimation. In *CVPR Workshops*, pages 2308–2317, June 2022. 2, 4

[32] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and

mesh estimation from a single rgb image. In *ECCV*, pages 752–768. Springer, 2020. 1, 5

[33] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, pages 440–455. Springer, 2020. 1

[34] 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 *CVPR*, pages 5079–5088, June 2018. 1, 2

[35] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, pages 548–564. Springer, 2020. 3, 5, 6, 7, 8

[36] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, pages 49–59, 2018. 1

[37] 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. *TOG*, 38(4):1–13, 2019. 1, 3, 5

[38] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, pages 1862–1869. IEEE, 2012. 1, 2

[39] Lingteng Qiu, Xuanye Zhang, Yanran Li, Guanbin Li, Xiaojun Wu, Zixiang Xiong, Xiaoguang Han, and Shuguang Cui. Peeking into occluded joints: A novel framework for crowd pose estimation. In *ECCV*, pages 488–504. Springer, 2020. 4

[40] Pengfei Ren, Haifeng Sun, Jiachang Hao, Jingyu Wang, Qi Qi, and Jianxin Liao. Mining multi-view information: a strong self-supervised framework for depth-based 3d hand pose and mesh estimation. In *CVPR*, pages 20555–20565, 2022. 1, 2

[41] Pengfei Ren, Haifeng Sun, Qi Qi, Jingyu Wang, and Weiting Huang. Srn: Stacked regression network for real-time 3d hand pose estimation. In *BMVC*, page 112, 2019. 5

[42] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *TOG*, 36(6), 2017. 3

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

[44] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9869–9878, 2020. 2, 6, 9

[45] Breannan Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, Jessica K Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *TOG*, 39(6):1–14, 2020. 1, 3

[46] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228. Springer, 2020. 1

[47] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, pages 89–98, 2018. 2

[48] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *ICCV*, pages 11698–11707, 2021. 2, 3, 4, 5

[49] 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. *TOG*, 36(6):1–12, 2017. 1, 3

[50] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *TOG*, 33(5):169:1–169:10, 2014. 1, 2

[51] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 118(2):172–193, 2016. 2, 6, 9

[52] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *CVPR*, pages 10853–10862, 2019. 1, 2

[53] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dense 3d regression for hand pose estimation. In *CVPR*, pages 5147–5156, 2018. 1, 2

[54] 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. *TOG*, 39(6):1–16, 2020. 2, 3, 5, 6, 9

[55] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, pages 52–67, 2018. 3, 4, 5

[56] Chao Wen, Yinda Zhang, Zhuwen Li, and Yanwei Fu. Pixel2mesh++: Multi-view 3d mesh generation via deformation. In *ICCV*, pages 1042–1051, 2019. 3

[57] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *CVPR*, pages 11354–11363, 2021. 2, 3, 7, 8

[58] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *ICCV*, pages 11446–11456, 2021. 3, 4, 5, 6

[59] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 1, 2

[60] Long Zhao, Xi Peng, Yu Tian, Mubbasir Kapadia, and Dimitris N Metaxas. Semantic graph convolutional networks for 3d human pose regression. In *CVPR*, pages 3425–3435, 2019. 2, 5

[61] Ce Zheng, Sijie Zhu, Matias Mendieta, Taojiannan Yang, Chen Chen, and Zhengming Ding. 3d human pose estima-

tion with spatial and temporal transformers. In *ICCV*, pages 11656–11665, 2021. 2, 5

[62] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, pages 5346–5355, 2020. 2

[63] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 1, 2

[64] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 1