MeMaHand: Exploiting Mesh-Mano Interaction for Single Image Two-Hand Reconstruction

Congyi Wang* Feida Zhu* Shilei Wen†
ByteDance
{kongyi.1990, zhufeida, zhengmin.666}@bytedance.com

Abstract

Existing methods proposed for hand reconstruction tasks usually parameterize a generic 3D hand model or predict hand mesh positions directly. The parametric representations consisting of hand shapes and rotational poses are more stable, while the non-parametric methods can predict more accurate mesh positions. In this paper, we propose to reconstruct meshes and estimate MANO parameters of two hands from a single RGB image simultaneously to utilize the merits of two kinds of hand representations. To fulfill this target, we propose novel Mesh-Mano interaction blocks (MMIBs), which take mesh vertices positions and MANO parameters as two kinds of query tokens. MMIB consists of one graph residual block to aggregate local information and two transformer encoders to model long-range dependencies. The transformer encoders are equipped with different asymmetric attention masks to model the intra-hand and inter-hand attention, respectively. Moreover, we introduce the mesh alignment refinement module to further enhance the mesh-image alignment. Extensive experiments on the InterHand2.6M benchmark demonstrate promising results over the state-of-the-art hand reconstruction methods.

1. Introduction

Vision-based 3D hand analysis plays an important role in many applications such as virtual reality (VR) and augmented reality (AR). Two-hand reconstruction from a single RGB image is more challenging due to complex mutual interactions and occlusions. Besides, the skin appearance similarity makes it difficult for the network to align image features to the corresponding hand.

Previous hand reconstruction works can be divided into two categories, parametric methods [3, 7, 32, 33] and non-parametric methods [4–6, 10, 18–20]. Parametric methods typically learn to regress *pose* and *shape* parameters of MANO model [23], where *pose* represents joint rotations in axis-angle representation and *shape* represents the coefficients of shape PCA bases. The MANO prior can yield plausible hand shapes from a single monocular image. However, they can not produce fine-grained hand meshes due to their limited capacity.

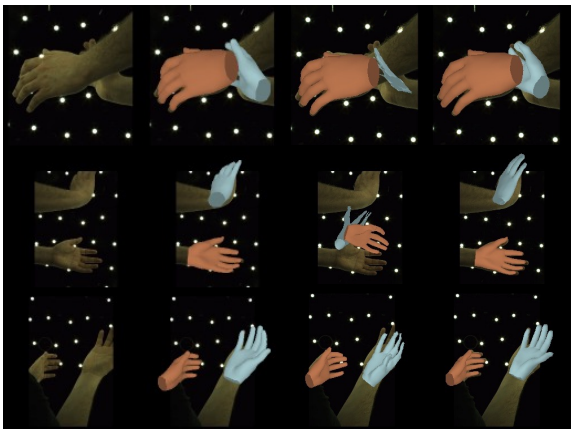


Figure 1. Comparison with the state-of-the-art method IntagHand [16] for single-image two-hand reconstruction. The integration of parametric and non-parametric hand representations allows us to achieve better performance in hard cases such as severe occlusions and challenging viewpoints.

With the rapid progress of graph convolutional network (GCN) and transformer techniques [10, 16, 18, 19], it is observed that direct mesh reconstruction can achieve state-of-the-art performance towards the Euclidean distances between the ground truth vertices and the predicted vertices. Nonetheless, the non-parametric methods are less robust in handling challenging viewpoints or severe occlusions.

In this paper, we introduce a novel single-image two-hand reconstruction method designed to predict mesh vertices positions and estimate MANO parameters simultaneously to utilize the merits of two kinds of hand representations. The proposed **Me**sh-**Ma**no interaction **Hand** reconstruction architecture (MeMaHand) consists of three modules: 1) the image encoder-decoder module, 2) the

* Equal contribution. † Corresponding author.

mesh-mano interaction module, 3) and the mesh alignment refinement module. To extract contextually meaningful image features, we pre-train a classical image encoder-decoder network on auxiliary tasks including hand segmentation, hand 2D joints and dense mapping encodings. The low-resolution features encode more global knowledge, while the high-resolution features contain more local details. Secondly, the mesh-mano interaction module stacks three mesh-mano interaction blocks (MMIBs) to transform the mesh vertices and MANO parameters queries initialized by the global image feature vector. We observe that the hand prior embedded in the MANO parameters is valuable for predicting stable hand meshes in challenging situations such as severe occlusions. MMIB consists of one graph residual block to aggregate local information and two transformer encoders to model long-range dependencies. The transformer encoders are equipped with different asymmetric attention masks to model the intra-hand and inter-hand attention, respectively. Each MMIB is followed by an upsampling operation to upsample the mesh vertices tokens in a coarse-to-fine manner. Finally, the mesh alignment refinement module utilizes one MMIB to predict offsets for mesh vertices and MANO parameters to enhance mesh-image alignment. To improve the reliability of image evidence, we project mesh vertices predicted by the Mesh-Mano interaction module onto the 2D image plane. The explicit mesh-aligned image features are concatenated to the transformer input tokens.

The whole network, including the pre-trained image feature encoder-decoder, is jointly optimized such that the image features better adapt to our hand mesh reconstruction task. Benefiting from the mesh-mano interaction mechanism and mesh alignment refinement stage, extensive experiments demonstrate that our method outperforms existing both parametric and non-parametric methods on InterHand2.6M [22] dataset. In summary, the contributions of our approach are as follows:

- We propose MeMaHand to integrate the merits of parametric and non-parametric hand representation. The mesh vertices and MANO parameters are mutually reinforced to achieve better performance in mesh recovery and parameter regression.
- A mesh-image alignment feedback loop is utilized to improve the reliability of image evidence. Therefore, more accurate predictions are obtained by rectifying the mesh-image misalignment.
- Our method achieves superior performance on the InterHand2.6M dataset, compared with both non-parametric and parametric methods.

2. Related Works

Parametric Hand Reconstruction. Parametric approaches [1–3, 7, 11, 12, 23, 32, 33] use a parametric hand model such as MANO [23] and focus on regressing the *pose* and *shape* parameters from a single image. The rich embedded prior information (*e.g.* the geometric dynamic constraints of joint rotations) can assist deep model learning when 3D annotations are insufficient. Weak 2D joints supervision [3] and motion caption data [33] are utilized to train convolutional neural networks (CNN) to predict MANO parameters. The reliance on 3D manual annotations is further totally alleviated by S^2HAND [7]. However, the reconstructed mesh can not fully express the local details of variable 3D hand shapes due to the limited capacity of the MANO model.

Non-parametric Hand Reconstruction. Non-parametric methods aim to reconstruct the hand mesh directly [5, 6, 8, 10, 16–21, 26, 29]. To explicitly encode mesh topology, graph convolutional network (GCN) is adopted for aggregating adjacent vertices features. Hierarchical architectures [10, 16] are designed for mesh generation from coarse to fine. To model long-range dependencies, multi-layer transformer encoders are introduced such that the global interactions can be modeled without being limited by any mesh topology [18, 19]. Most recently, IntagHand [16] achieves the state-of-the-art performance on InterHand2.6M dataset [22]. However, we observe that IntagHand is less robust in handling challenging viewpoints or severe occlusions. The reconstructed mesh may be corrupted into unnatural shapes.

Interacting Two-Hand Reconstruction. Although single-hand methods can extend to two-hand reconstruction, the correlation between the left and right hands is not considered. Besides, the performance deteriorates for the close-interacting hands. Depth cameras [15] are sensitive to tracking accuracy and multi-view images [24, 27] are expensive to acquire. Based on a large-scale interacting hand dataset named InterHand2.6M [22], deep learning based single image methods either estimate hand joint positions [14, 22], or employ a 2.5D heatmap to predict the MANO parameters [30], or directly reconstruct meshes [16]. In summary, existing methods only employ a single specific hand representation. In contrast, our method not only combines the merits of MANO representations and mesh representations but also proposes a dedicated transformer encoder with asymmetric attention masks to make them collaborate seamlessly. Besides, the joint rotation and precise mesh vertices positions can be used for different applications, such as driving CG characters or virtual try-on.

3. Methodology

The overall architecture of MeMaHand is depicted in Fig. 2. Given a single image I, the proposed method can predict the mesh vertices positions $V \in \mathbb{R}^{778\times3}$, MANO pose parameter $\theta \in \mathbb{R}^{48}$ and shape parameter $\beta \in \mathbb{R}^{10}$ of two hands simultaneously in one forward pass. In this section, we first describe the overall architecture of MeMa-

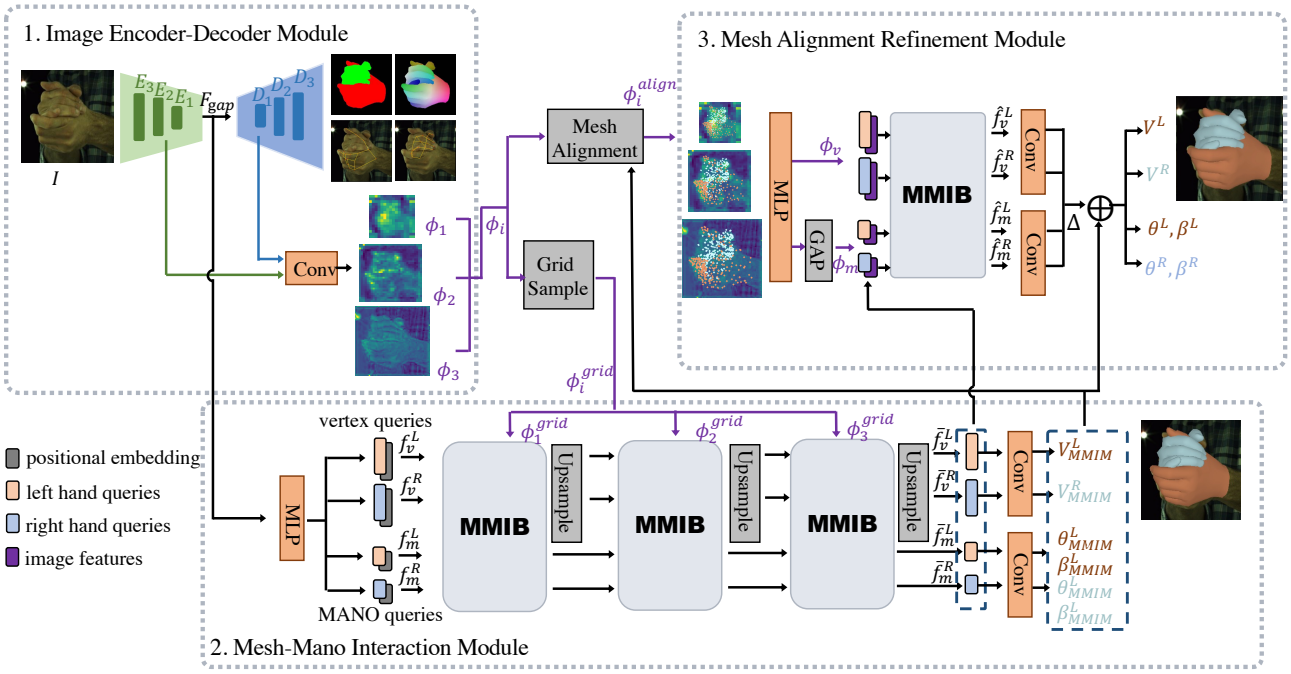


Figure 2. The overall architecture of our proposed MeMaHand. We pretrain an image encoder-decoder module on auxiliary tasks to extract multi-scale contextually meaningful image features. The Mesh-Mano interaction module stacks three Mesh-Mano interaction blocks (MMIBs) to predict the mesh vertices positions and MANO parameters simultaneously. The mesh alignment refinement module rectifies the mesh-image misalignment by explicitly utilizing the mesh-align image features.

Hand in Sec. 3.1. Then, we illustrate the mesh-mano interaction module and mesh-alignment refinement module in Sec. 3.2 and Sec. 3.3, respectively. Finally, we introduce the model objectives used to train our network in Sec. 3.4.

3.1. System Overview

To extract contextually meaningful image features, we pre-train a classical image encoder-decoder module on auxiliary tasks including hand segmentation, hand 2D joints, and dense mapping encodings. The 2D image conditions are generated from the reconstructed mesh of ground-truth MANO parameters. ResNet50 [13] is leveraged as the backbone. Global average pooling feature F_{gap} and multi-scaled image feature $\{\phi_i \in \mathbb{R}^{C_i \times H_i \times W_i}, i = 1, 2, 3\}$ are extracted from the image encoder (E) and decoder (D).

Afterward, we propose Mesh-Mano Interaction Module to predict mesh vertices positions V_{MMIM} and MANO parameters $\theta_{MMIM}, \beta_{MMIM}$ of both left (L) and right (R) hands. We use the global feature F_{gap} to initialize the vertex and MANO queries. The grid-sampled image features $\{\phi_i^{grid}, i = 1, 2, 3\}$ are tokenized as well following the practice of Mesh Graphformer [19].

The mesh alignment refinement module further improves the predictions generated by the mesh-mano interaction module. The explicit mesh-aligned image features extracted from the multi-scale feature maps ϕ_i are more informative than grid-sampled image features, which helps rectify the predictions. Details will be elaborated in Sec. 3.3.

3.2. Mesh-Mano Interaction Module

The mesh-mano interaction module reconstructs hand mesh in a coarse-to-fine manner with three Mesh-Mano Interaction Blocks (MMIBs). We leverage the graph coarsening method [16] to build three-level sub meshes with vertex number $N_1 = 63, N_2 = 126, N_3 = 252$. Each MMIB is followed by an upsample operation that reverses the topological relationship between adjacent sub-meshes. The full mesh vertices positions ($N = 778$) are obtained with a simple 1×1 $Conv$ from the final output vertex token.

Fig. 3 shows the detailed structure of MMIB. At level i, the input tokens of MMIB include two kinds of queries: vertex queries $f_v^h \in \mathbb{R}^{N_i \times D_i}$ and MANO parameter queries $f_m^h \in \mathbb{R}^{2 \times D_i}$, where h indicates left (L) or right (R) hand, and D_i represents the feature dimensions. Therefore, the total sequence length is $(2 \times N_i + 4)$.

Graph Residual Block. The design of the graph residual block is similar to [10, 16]. The Chebyshev spectral graph CNN [9] is adopted to transform the vertex token to intermediate graph features,

$$f_{graph} = \mathbf{GraphConv}(f_v^L, f_v^R), \tag{1}$$

where the operation **GraphConv** denotes the graph con-

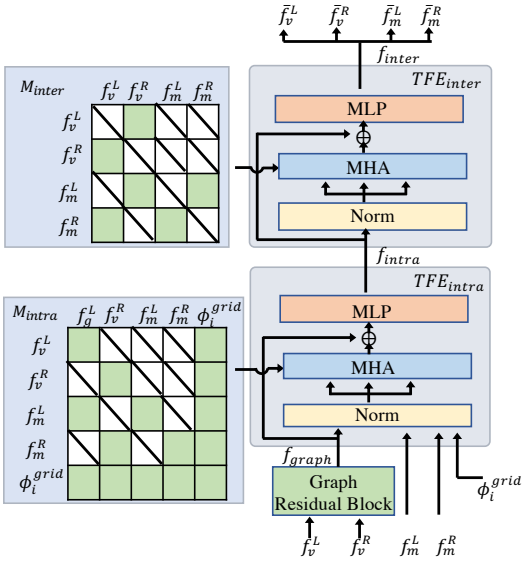


Figure 3. The detailed structure of Mesh-Mano Interaction Block (MMIB). MMIB consists of one graph residual block to aggregate local information and two transformer encoders to model long-range dependencies.

volution. More details can be found in [9]. For simplicity, we use f_{graph} to represent the concatenation of updated left-hand and right-hand vertex tokens.

Intra-hand Transformer Encoder. While graph CNN is useful for extracting local information, it is less efficient at capturing long-range dependencies. We use transformer encoders to model dependencies not only between long-range vertices but also between mesh vertices and MANO parameters. The spatial image features at resolution i are also tokenized as ϕ_i^{grid} by grid sampling following the practice of Mesh Graphformer [19]. Therefore, the input tokens of intra-hand transformer encoder (TFE_{intra}) consist of graph features (f_{graph}), MANO queries (f_m^L, f_m^R) and grid image feature (ϕ_i^{grid}). Based on the observation that the mesh vertices regression can lead to more precise mesh reconstruction and the MANO parameters are more stable, we propose an asymmetric attention mask excluding the mano-to-mesh attention such that the MANO parameter tokens will not directly affect the mesh vertex tokens. The mesh-to-mano attention remains. Therefore, the mesh vertex tokens are implicitly regularized by the MANO parameters loss. TFE_{intra} focuses on modeling the intra-hand dependencies. The inter-hand attention is also excluded. The resulting asymmetric attention mask M_{intra} is shown in Fig. 3. Finally, the intra-hand transformer encoder is formulated as:

$$f_{intra} = TFE_{intra}([f_{graph}, f_m^L, f_m^R, \phi_i^{grid}], M_{intra}), \tag{2}$$

where $[...]$ denotes the concatenation operation. For simplicity, f_{intra} represents the concatenation of updated mesh vertex and MANO parameter tokens.

Inter-hand Transformer Encoder. In contrast to TFE_{intra}, the inter-hand transformer encoder (TFE_{inter}) focuses on modeling the inter-hand correlations. Fig. 3 presents the resulting asymmetric attention mask M_{inter}. The image feature tokens are not used in TFE_{inter}. Finally, the inter-hand transformer encoder is formulated as:

$$f_{inter} = TFE_{inter}(f_{intra}, M_{inter}), \tag{3}$$

where f_{inter} can be further split into updated vertex tokens $\bar{f}_v^L, \bar{f}_v^R$ and MANO parameter tokens $\bar{f}_m^L, \bar{f}_m^R$.

Our proposed MMIB predicts the mesh vertices positions and MANO parameters simultaneously in a unified architecture. Intra-hand and inter-hand mesh-mano dependencies are modeled by two cascaded transformer encoders with different attention masks. After three MMIBs, a simple 1×1 $Conv$ is applied to obtain the mesh vertices positions and MANO parameters from output tokens,

$$V_{MMIM}^h = Conv_v(\bar{f}_v^h), \tag{4}$$

$$\theta_{MMIM}^h, \beta_{MMIM}^h = Conv_m(\bar{f}_m^h), \tag{5}$$

where $h \in \{L, R\}$ indicates the left or right hand. To tackle the noticeable misalignment between the estimated meshes and image evidence, we propose a mesh alignment refinement module to rectify the results.

3.3. Mesh Alignment Refinement Module

To further improve mesh-image alignment, we propose a novel mesh-alignment refinement module inspired by PyMAF [31]. Specifically, we project the mesh vertices V_{MMIM} predicted by the mesh-mano interaction module onto the multi-scale image features explicitly. Three-scale mesh-aligned image features are concatenated together for each corresponding vertex token. A simple multi-layer perceptron (MLP) is adopted to reduce the dimensions, resulting in a fused image feature ϕ_v.

We utilize one MMIB to refine the mesh vertex and MANO parameter tokens $\bar{f}_v^h, \bar{f}_m^h$. To effectively utilize the mesh-aligned image evidence, we made some modifications to the tokens. Specifically, we concatenate the image feature ϕ_v to vertex token $\bar{f}_v^h$ along the channel dimension. For MANO parameter tokens, we perform a global average pooling operation to obtain global feature vector ϕ_m before concatenation. The rectified tokens are formulated as,

$$\hat{f}_v^h, \hat{f}_m^h = MMIB(\bar{f}_v^h, \bar{f}_m^h, \phi_v, \phi_m). \tag{6}$$

One simple 1×1 $Conv$ layer is applied to obtain the mesh vertices positions offsets and MANO parameter off-

sets for rectification,

$$V^h = V^h_{MMIM} + Conv_v(\hat{f}^h_v), \tag{7}$$

$$\theta^h = \theta^h_{MMIM} + Conv_m(\hat{f}^h_m), \tag{8}$$

$$\beta^h = \beta^h_{MMIM} + Conv_m(\hat{f}^h_m), \tag{9}$$

where V^h, θ^h and β^h are the final results. Note that we can stack several MMIBs to predict the offsets iteratively. In our experiments, one single MMIB is enough to achieve satisfactory results. PyMAF [31] fuses the mesh-align features into one global feature, which ignores the inherent spatial relationship of mesh vertices. In contrast, our MMIB adopts GCN to model the spatial relation between adjacent mesh vertices. The transformer encoder is responsible for modeling the long-range dependencies.

3.4. Model Objectives

Our method predicts mesh vertices positions and MANO parameters simultaneously. The learning objectives can be divided into three categories: mesh vertex loss, MANO parameter loss and mesh-mano consistency loss.

Mesh Vertex Loss: The widely-used L1 loss is adopted to supervise the vertex positions and the 2D projections:

$$\mathcal{L}_V = \sum_{h=L,R} \|V^h - V^h_{GT}\|_1 + \|\Pi(V_h) - \Pi(V^h_{GT})\|_1, \tag{10}$$

where V^h_{GT} represents the ground-truth vertex positions and Π denotes the projection function. Given vertex positions, the joint positions can be regressed by multiplying the predefined regression matrix $\mathcal{J}$. Joint losses are formulated as:

$$\begin{aligned}\mathcal{L}_J = &\sum_{h=L,R} \|\mathcal{J}V^h - \mathcal{J}V^h_{GT}\|_1 \\ &+ \sum_{h=L,R} \|\Pi\mathcal{J}(V_h) - \Pi(\mathcal{J}V^h_{GT})\|_1.\end{aligned} \tag{11}$$

The face normal loss is introduced to regularize the surface normal consistency:

$$\mathcal{L}_N = \sum_{h=L,R} \sum_{f=1}^{F} \sum_{i=1}^{3} \|e^h_{f,i} \cdot n^h_{f,GT}\|_1, \tag{12}$$

where $e^h_{f,i}$ represents the ith edge of face f at hand h and $n^h_{f,GT}$ is the normal vector of this face from the ground truth mesh. The edge length consistency loss is to enforce the edge length consistency:

$$\mathcal{L}_E = \sum_{h=L,R} \sum_{i=1}^{E} \|e^h_i - e^h_{i,GT}\|_1, \tag{13}$$

where e^h_i represents the ith edge of hand h and E denotes the total edge number, respectively.

Method	Params(M)	Inference Time (ms)
InterShape [30]	139.6	47.77
Intaghand [16]	37.28	51.86
Ours	38.31	60.52

Table 1. Parameters and inference speed on Tesla V100

MANO Parameter Loss: Given the ground truth MANO parameters, L1 loss is used to regress parameters,

$$\mathcal{L}_P = \sum_{h=L,R} \|\theta^h - \theta^h_{GT}\|_1 + \|\beta^h - \beta^h_{GT}\|_1. \tag{14}$$

In addition, we reconstruct hand mesh based on the MANO model. The reconstructed mesh should be close to the ground-truth mesh,

$$\mathcal{L}_{V,mano} = \sum_{h=L,R} \|\mathcal{MANO}(\theta^h, \beta^h) - V^h_{GT}\|_1. \tag{15}$$

Mesh-Mano Consistency Loss. The predicted mesh and the reconstructed mesh from MANO parameters should be consistent with each other:

$$\begin{aligned}\mathcal{L}_{consist} = &\sum_{h=L,R} \|\mathcal{MANO}(\theta^h, \beta^h) - V^h\|_1 \\ &+ \|\mathcal{JMANO}(\theta^h, \beta^h) - \mathcal{J}V^h\|_1.\end{aligned} \tag{16}$$

In summary, the overall training loss $\mathcal{L}$ is:

$$\begin{aligned}\mathcal{L} = &\lambda_V\mathcal{L}_V + \lambda_J\mathcal{L}_J + \lambda_N\mathcal{L}_N + \lambda_E\mathcal{L}_E \\ &+ \lambda_P\mathcal{L}_P + \lambda_{V,mano}\mathcal{L}_{V,mano} + \lambda_{consist}\mathcal{L}_{consist}.\end{aligned} \tag{17}$$

where $\lambda_V = 40$, $\lambda_J = 40$, $\lambda_N = 5$, $\lambda_E = 40$, $\lambda_P = 10$, $\lambda_{V,mano} = 10$ and $\lambda_{consist} = 40$. The whole network including the pre-trained image feature encoder-decoder is jointly optimized such that the image features better adapt to our hand mesh reconstruction task.

4. Experiment

4.1. Datasets and Implementation

Training Dataset. Interhand2.6M [22] is a large-scale hand dataset with ground truth mesh annotations including both single-hand and interacting two-hand images. We pick out the interacting hand (IH) data for training and testing. For a fair comparison, we follow the preprocessing steps of [16] which produces 366K training and 261K testing samples.

Implementation Details. The image encoder uses ResNet50 [13] as the backbone. The decoder contains four simple deconvolutional layers. We first pretrain our image encoder-decoder module on auxiliary tasks. Then, the whole network is jointly optimized with Adam optimizer at

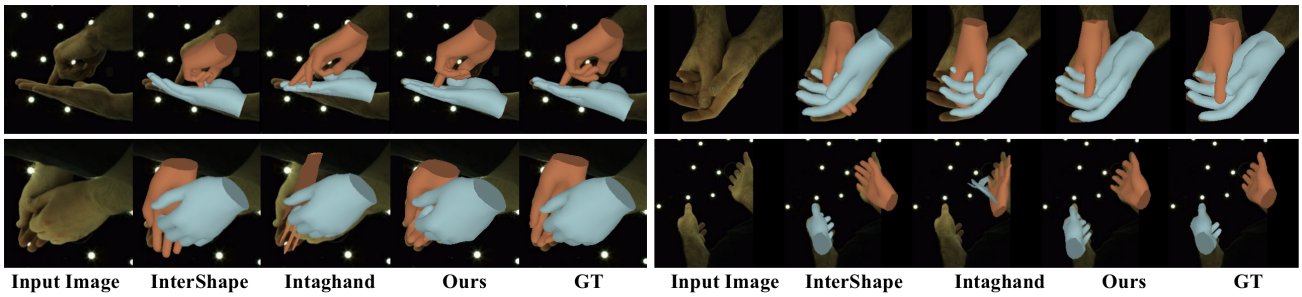


Figure 4. Qualitative comparison with SOTA parametric method InterShape [30] and SOTA non-parametric method IntagHand [16] on InterHand2.6M dataset. Our method performs better on close-interacting two-hand reconstruction (first row). Besides, our method is robust in hard cases such as severe occlusions and challenging viewpoints (second row).

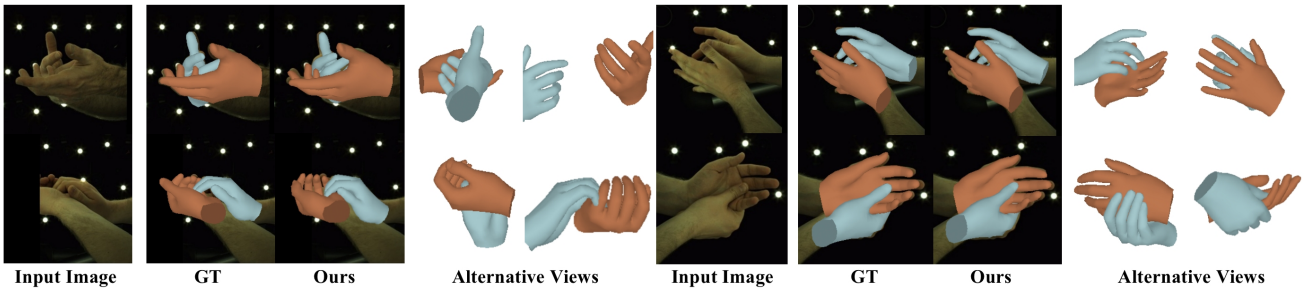


Figure 5. Our method produces accurate hand reconstructions for various two-hand interactions.

Method	MPJPE↓	MPVPE↓	AUC ↑	PROJ2D↓
zimmermann *et al.* [34]	36.36	-	-	-
Spurr *et al.* [25]	15.40	-	-	-
InterNet [22]	16.00	-	0.711	-
Intaghand [16]	8.79	9.03	0.806	6.47
Ours	**8.65**	**8.89**	**0.832**	**6.22**

Table 2. Quantitative comparison with state-of-art non-parametric methods on InterHand2.6M dataset.

a learning rate 1×10^{-4}. Data augmentation includes random rotation, random translation, and random scaling. The batch size is set to 32. It takes around 48 hours to train MeMaHand with 4 Tesla V100 GPUs.

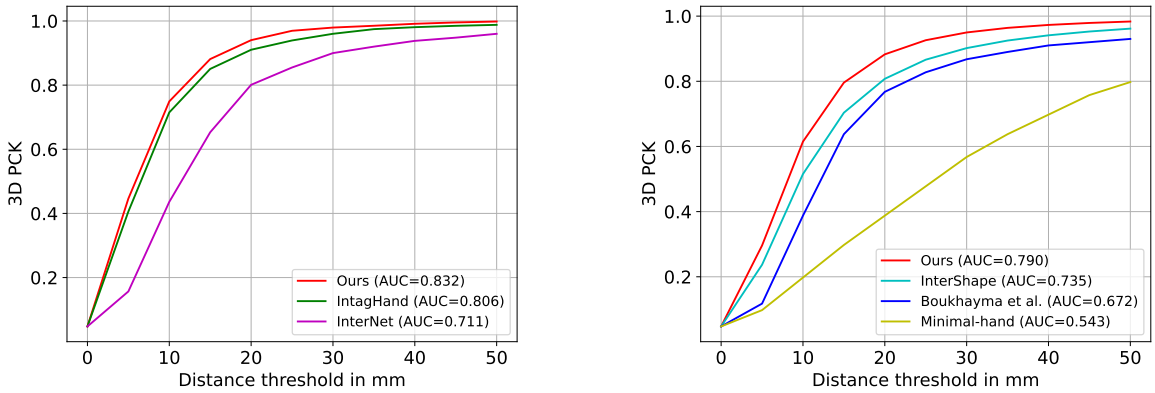
Metrics. Mean Per Joint Position Error (MPJPE) and Mean Per Vertex Position Error (MPVPE) in millimeters are adopted to evaluate the mesh reconstruction accuracy. Additionally, we report the percentage of correct keypoints (PCK) curve and Area Under the Curve (AUC) across the thresholds between 0 and 50 millimeters. To evaluate the mesh-image alignment accuracy, the reconstructed mesh vertices are projected onto the 2D image plane. The mesh-image alignment accuracy is measured by PROJ2D, which calculates the distance in image pixels between the projected ground truth vertices and the predicted vertices.

4.2. Comparison with State-of-the-art Methods

We compare our model with state-of-the-art non-parametric methods including InterNet [22] and Intaghand [16] and parametric methods including [3], MinimalHand [33] and InterShape [30]. The officially released weights are used to obtain the results. The model parameters and inference time of our method and other SOTA methods are reported in Tab. 1. The inference of our model can be completed in $60.52ms$, comparable to other SOTAs but with better mesh reconstruction results. Besides, the mesh vertex positions as well as the MANO parameters are predicted simultaneously in one single forward pass.

Comparison with Non-parametric Methods. Non-parametric methods directly generate mesh vertice positions, which can express the local details of variable 3D hand shapes. The quantitative results of non-parametric methods are shown in Tab. 2. It can be seen that our method achieves the best performance on all evaluation metrics. Fig. 6(a) presents the PCK curve, which further demonstrates the superiority of our method at all threshold levels.

Fig. 4 presents the qualitative comparison with SOTA non-parametric method IntagHand [16]. Our method performs better on close interacting two-hand reconstruction (first row). Besides, we observe that IntagHand produces collapsed meshes in hard cases such as severe occlusions



(a) Comparison with non-parametric method

(b) Comparison with parametric method

Figure 6. Comparison with the state-of-the-art methods on InterHand2.6M dataset. **Left**: 3D PCK of non-parametric methods. **Right**: 3D PCK of parametric methods. The AUC values are shown in parentheses.

Method	MPJPE↓	MPVPE ↓	AUC↑	PROJ2D↓
Boukhayma *et al*. [3]	16.93	17.98	0.672	-
MinimalHand [33]	23.48	23.89	0.543	-
InterShape [30]	13.48	13.95	0.735	9.95
Ours	**10.85**	**10.92**	**0.790**	**6.90**

Table 3. Quantitative comparison with state-of-art parametric methods on InterHand2.6M dataset.

and challenging viewpoints (second row). In contrast, our method is more robust to such situations. We attribute this success to the integration of MANO representations, which predicts stable hand meshes. More results with alternative views are shown in Figure 5.

Comparison with Parametric Methods. For a fair comparison, the reconstructed hand meshes from our predicted MANO parameters are utilized for evaluation. The quantitative comparisons are listed in Tab. 3. The PCK curve is shown in Fig. 6(a). The MANO parameter estimation of our method outperforms other SOTA parametric methods. Thanks to the mesh-mano interaction block, the MANO token is conditioned on the mesh vertex tokens where spatial relations are retained. Such dependencies are more informative than one-dimensional global feature utilized in competing methods [3, 30, 33].

From Tab. 2 and Tab. 3, we can find the parametric method InterShape [30] has higher statistical errors on joint and vertex positions compared with the non-parametric method IntagHand [16] due to the limited capacity of the parametric model. However, from the second row of Fig. 4, we can see InterShape produces reasonable hand shapes in these challenging cases. By combining the merits of parametric and non-parametric hand representations, our method achieves the best performance both quantitatively and qualitatively.

Figure 7. Two-hand reconstruction of images in the wild, which are taken from the RGB2Hands [28].

4.3. Extension to Images in the Wild

We further present the hand reconstruction results on images in the wild. As shown in Fig. 7, our method performs well on real-life images taken from the RGB2Hands dataset [28], which demonstrates the generalization ability of our approach. Recall that our method is designed to predict accurate mesh vertices positions and MANO parameters simultaneously in one forward pass. The MANO *pose* parameter represents the joint rotations in axis-angle representation, which is useful for animating 3D hands in computer graphics. Accurate hand mesh vertices positions and MANO parameters for hands in the wild can facilitate different human-computer-interaction (HCI) applications.

4.4. Ablation Study

To evaluate the effectiveness of the proposed modules in our framework, we conduct an ablation study on several variants of our method. The quantitative results of all vari-

	MPJPE↓	MPVPE↓	AUC↑	PROJ2D↓
A: w/o MANO token	8.87	9.09	0.818	6.85
B: w/o Mesh-Align	8.79	9.01	0.826	6.64
C: Mesh-Align scale-16	8.73	8.97	0.829	6.38
D: Mesh-Align scale-64	8.77	8.99	0.826	6.58
E: Regress MANO from F_{gap}	8.83	9.04	0.828	6.86
F: w/o asymmetric attention	9.17	9.37	0.822	6.67
G: w/o auxiliary tasks	9.11	9.28	0.823	6.78
Full Model	**8.65**	**8.89**	**0.832**	**6.22**

Table 4. Comparisons of different variants of our method.

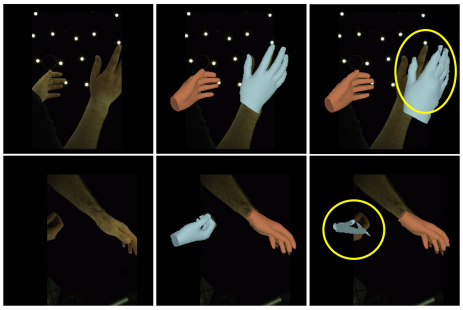


Figure 8. Ablation study on MANO tokens.

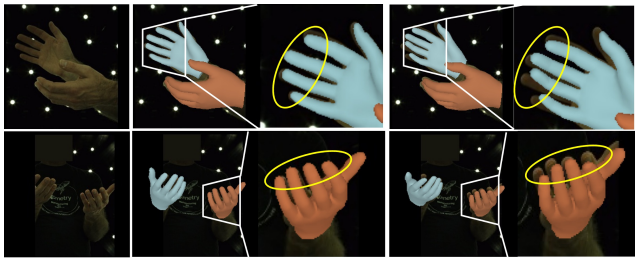


Figure 9. Ablation study on mesh alignment refinement.

ants are presented in Table 4.

Effectiveness of MANO token. Variant **A** (w/o MANO token) represents removing the MANO tokens. The transformer encoder only models the dependencies between the mesh vertex tokens and image features. Fig. 8 show the qualitative comparison. Without hand prior information, the reconstructed mesh may be corrupted when parts of the hands are occluded. In contrast, our full model can generate reasonable meshes in these challenging situations.

Effectiveness of Mesh Alignment Refinement. Variant **B** (w/o Mesh-Align) represents removing the mesh alignment refinement module. The predictions generated by the Mesh-Mano interaction module are taken as the final outputs. Variant **C** and **D** denote extracting the single-scale image features at resolution 16 and 64, respectively. The performances degrade without using mesh align refinement or using single-scale image features. As shown in Fig. 9, our full model produces better-aligned results.

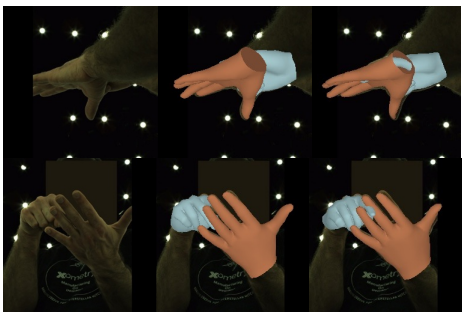


Figure 10. Failures cases caused by severe occlusions.

Effectiveness of MMIB. Parametric and non-parametric representations are widely used for many years. However, combining them is not trivial. In variant **E** (Regress MANO from F_{gap}), MMIB is responsible for updating vertex tokens. The MANO parameters are regressed from the global average pooling features F_{gap} with simple fully connected layers rather than through MMIB. The performance gain of variant **E** is limited as shown in Tab. 4. On the other hand, we propose an asymmetric attention mask to exclude the mano-to-mesh attention. To verify the choice, variant **F** (w/o asymmetric attention mask) denotes removing the asymmetric attention mechanism. The performance degrades compared with our full model.

Effectiveness of Pretraining on Auxiliary Tasks. Variant **G** represents the backbone is pre-trained on ImageNet. Our full model performs better since pre-training on auxiliary tasks are essential to extract semantically-meaningful image features. The auxiliary 2D image conditions are generated from the ground-truth mesh of Interhand2.6M training split. We did not use extra datasets.

5. Conclusion

In this paper, we propose a novel approach MeMaHand for two-hand mesh reconstruction from a single image. The mesh-mano interaction module combines the merits of non-parametric and parametric representations. Then, the mesh alignment refinement module further rectifies the results with an explicit mesh alignment feedback loop. Extensive experiments on the InterHand2.6M benchmark demonstrate that our proposed MeMaHand is superior to both existing parametric and non-parametric methods.

Limitations. Although our method generates promising results, it still fails in cases of severe occlusions. Fig. 10 shows some failure cases, where inter-penetration occurs. Taking the physical plausibility into consideration will be our future work.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020. 2

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 1, 2, 6, 7

[4] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 1

[5] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12929–12938, 2021. 1, 2

[6] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13283, 2021. 1, 2

[7] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. 1, 2

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

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016. 3, 4

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 1, 2, 3

[11] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 2

[12] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3, 5

[14] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11189–11198, 2021. 2

[15] Nikolaos Kyriazis and Antonis Argyros. Scalable 3d tracking of multiple interacting objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3430–3437, 2014. 2

[16] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2761–2770, 2022. 1, 2, 3, 5, 6, 7

[17] Guan Ming Lim, Prayook Jatesiktat, and Wei Tech Ang. Mobilehand: Real-time 3d hand shape and pose estimation from color image. In *International Conference on Neural Information Processing*, pages 450–459. Springer, 2020. 2

[18] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1954–1963, 2021. 1, 2

[19] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. 1, 2, 3, 4

[20] 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 *European Conference on Computer Vision*, pages 752–768. Springer, 2020. 1, 2

[21] 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 *European Conference on Computer Vision (ECCV)*, 2020. 2

[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*, pages 548–564. Springer, 2020. 2, 5, 6

[23] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Graph.*, 36(6), nov 2017. 1, 2

[24] 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 (TOG)*, 39(6):1–14, 2020. 2

[25] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018. 6

[26] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11698–11707, October 2021. 2

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

[28] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 7

[29] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 52–67, 2018. 2

[30] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021. 2, 5, 6, 7

[31] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11446–11456, 2021. 4, 5

[32] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 1, 2

[33] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 1, 2, 6, 7

[34] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 6