# Overcoming the Trade-off Between Accuracy and Plausibility in 3D Hand Shape Reconstruction

Ziwei Yu[1] Chen Li[1] Linlin Yang[1] Xiaoxu Zheng[2] Michael Bi Mi[2] Gim Hee Lee[1] Angela Yao[1]
[1]National University of Singapore [2]Huawei International Pte Ltd, Singapore
{yuziwei, lichen}@u.nus.edu, {mu4yang, zhengxiaoxu66}@gmail.com, michaelbimi@yahoo.com
{gimhee.lee, ayao}@comp.nus.edu.sg

## Abstract

*Direct mesh fitting for 3D hand shape reconstruction is highly accurate. However, the reconstructed meshes are prone to artifacts and do not appear as plausible hand shapes. Conversely, parametric models like MANO ensure plausible hand shapes but are not as accurate as the non-parametric methods. In this work, we introduce a novel weakly-supervised hand shape estimation framework that integrates non-parametric mesh fitting with MANO model in an end-to-end fashion. Our joint model overcomes the tradeoff in accuracy and plausibility to yield well-aligned and high-quality 3D meshes, especially in challenging two-hand and hand-object interaction scenarios.*

## 1. Introduction

State-of-the-art monocular RGB-based 3D hand reconstruction methods [6, 10, 17, 21, 22, 28] focus on recovering highly accurate 3D hand meshes. As accuracy is measured by an average joint or vertex position error, recovered hand meshes may be well-aligned in 3D space but still be physically implausible. The 3D mesh surface may have irregular protrusions or collapsed regions (see Fig. 1), especially around the fingers. The meshes may also suffer from incorrect contacts or penetrations when there are hand-object or two-hand interactions. Yet methods that prioritize physical plausibility, especially in interaction settings [3,8,10,14,20,21], are significantly less accurate in 3D alignment. In summary, the current body of work predominantly favours either 3D alignment accuracy or physical plausibility, but cannot achieve both.

A closer examination reveals that the trade-off between 3D alignment and plausibility is also split methodology-wise. Methods that use the MANO model [30] produce plausible hand poses and hand shapes [2, 7, 38, 40, 42] due to the statistical parameterization of MANO. However, it is challenging to directly regress these parameters, since the mapping from an image to the MANO parameter space is highly non-linear. As a consequence, MANO-based methods lag in 3D alignment accuracy compared to non-parametric methods.

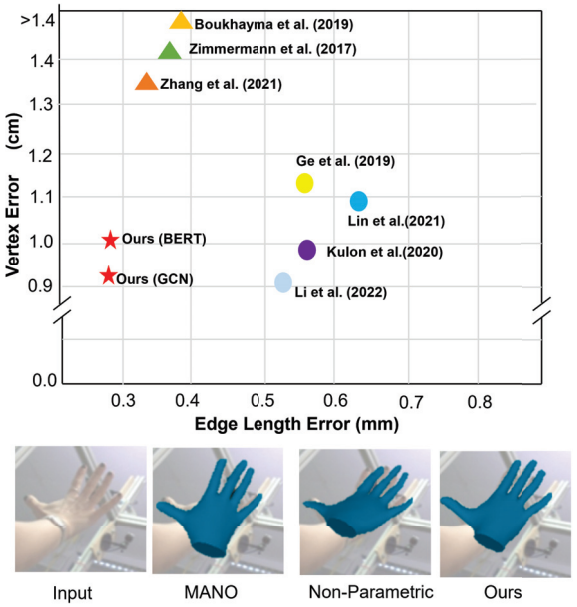


Figure 1. The vertex error vs. edge length error indicates that existing methods trade-off alignment accuracy with plausibility. MANO-based methods (triangles) vs. non-parametric model-based methods (circles) have a trade-off between vertex error and edge length error; our combined method (stars) can overcome this trade-off to yield well-aligned and plausible meshes. Plot of results from InterHand2.6M [27]; visualization from FreiHand [42].

Non-parametric methods [6, 10, 11, 17, 18, 21, 22, 28] directly fit a 3D mesh to image observations. Direct mesh fitting is accurate but is prone to surface artifacts. In scenarios with hand-object or hand-hand interactions, mesh penetrations cannot be resolved meaningfully even with regularizers such as contact losses [14] due to the unconstrained

optimization. Attention mechanism [21, 32] can mitigate some penetrations and artifacts, but the inherent problem remains. As such, the favoured approaches for hand-object and hand-hand interactions are still driven by MANO models [3, 8, 13, 14, 38].

In this work, we aim to recover high-quality hand meshes that are accurately aligned *and* have minimal artifacts and penetrations. To avoid a trade-off, we leverage direct mesh fitting for alignment accuracy and guidance from MANO for plausibility. Combine the non-parametric and parametric models is straightforward in terms of motivation. However, merging the two is non-trivial because it requires a mapping from non-parametric mesh vertices to parametric model parameters. This mapping, analogous to the mapping from an RGB image, is highly non-linear and difficult to achieve directly [16].

One of our key contributions in this work is a method to accurately map non-parametric hand mesh vertices to MANO joint parameters $\theta$. To do so, we perform a two-step mapping, from mesh-vertices to the joint coordinates, and joints coordinates to $\theta$. In the literature, the common practice for the former is to leverage the $\mathcal{J}$ matrix in MANO and linearly regress the joints from the mesh [17, 20, 21]. Yet the $\mathcal{J}$ matrix was designed to only map MANO-derived meshes to joints in a rest pose (see Eq. 10 in [25]). As we show in our experiments, applying $\mathcal{J}$ to non-rest poses introduces a gap of around 2 mm. Furthermore, we postulate that there is a domain gap between the estimated non-parametric meshes and MANO-derived meshes, even if both meshes have the same topology. To close this gap – we propose a VAE correction module, to be applied after the linear regression with $\mathcal{J}$. To map the recovered joints from the mesh to $\theta$, we use a twist-swing decomposition and analytically compute the $\theta$. It has been shown previously in [20] that decomposing joint rotations into twist-swing rotations [1] can simplify the estimation of human SMPL [25] model pose parameters. Inspired by [20], we also leverage the decomposition and further verify that the twist angle has minimal impact on the hand.

Note that obtaining ground truth labels for hand mesh vertices is non-trivial. Our framework lends itself well for weak-supervision. Since the estimated 3D mesh from the non-parametric decoder is regressed into 3D joints, it can also be supervised with 3D joints as weak labels (see Fig 2). At the same time, the parametric mesh estimated from these joints can be used as a pseudo-label for learning the non-parametric mesh vertices. Such a procedure distills the knowledge from the parametric model and is effective without ground truth mesh annotations. We name our method **WSIM 3D Hand**, in reference to **W**eakly-supervised **S**elf-distillation **I**ntegration **M**odel for **3D hand** shape reconstruction. Our contributions include:

- A novel framework that integrates a parametric and a non-parametric mesh model for accurate *and* plausible 3D hand reconstruction.
- A VAE correction module that closes the overlooked gap between non-parametric and MANO 3D poses;
- A weakly-supervised pipeline, competitive to a fully-supervised counterpart, using only 3D joint labels to learn 3D meshes.
- Significant improvements over state-of-the-art on hand-object or two-hand interaction benchmark datasets, especially in hand-object interaction on DexYCB.

## 2. Related Work

**Parametric Methods.** Previous works [2, 3, 7, 13, 14, 35, 36, 38–40] in 3D hand shape reconstruction often used the MANO model [30] to estimated 3D hand meshes. Boukhayma et al. [2] first use a deep neural network to regress the MANO parameters in single-hand reconstruction. However, directly estimating MANO parameters accurately is challenging, as they sit in an abstract PCA space. Moreover, previous MANO model-based methods ignore the spatial information that limited their reconstruction accuracy [5, 10]. This work addresses the above drawbacks of the MANO model by integrating a non-parametric model.

**Non-parametric Methods** Non-parametric 3D hand shape methods [6, 10, 12, 17, 21–23, 28, 32, 37] directly fit the mesh vertices either with graph convolutional networks [9, 11] or transformers [33]. Initially, spectral graph neural networks were used [9] but as they are not able to leverage deeper neighbourhood nodes' information, spatial graphs with spiral convolutions [17] were proposed instead. Subsequently, works applied mesh transformers [22] and other attention mechanisms [12, 23, 28] to facilitate interaction modelling. The estimated 3D pose and shapes of non-parametric methods are highly accurate; however, their many degrees of freedom also yield implausible 3D shapes with artifacts. This work integrates non-parametric methods with a MANO model to achieve both alignment accuracy and plausibility.

**Hand-Object and Two-Hand Interactions** Hand interactions add challenge to 3D shape reconstruction due to the additional occlusion from the interacting object or hand and possibility of surface collisions. For hand-object interaction, previous works [3, 13, 14, 35, 39] leverage the MANO model to ensure plausible hand shapes during the interaction modelling. Similarly, for two-hand interactions [27, 38], MANO has been applied to the left and right hand individually to simplify the two-hand reconstruction

into two single hand parameters estimation pipeline. By using MANO in the interaction setting, these above works are able to estimate plausible 3D hand shapes, though the alignment accuracy generally lags compared to non-parametric methods [6, 21, 32] that are less plausible. Different from the above methods, we first use the non-parametric model to learn the 3D joints and then convert these joints into accurate MANO parameters in the interaction setting. Therefore, our work overcomes the tradeoff between plausibility and accuracy.

## 3. Method

Fig. 2 shows an overview of our method. It has three components: the RGB encoder network, a non-parametric pose decoder (Sec. 3.2) and a parametric mesh reconstruction (Sec. 3.3) based on the MANO model (Sec. 3.1). The overall framework can be learned in a weakly-supervised manner with self-distillation (Sec. 3.5). For hand-object and hand-hand interactions, we add an interaction refinement module to reduce the penetration (Sec. 3.4).

### 3.1. Preliminaries

MANO [30] is a statistical 3D hand pose and shape model. It maps pose parameters $\theta \in R^{16\times3}$ and shape parameters $\beta \in R^{10}$ to a 3D hand mesh with model $M(\beta,\theta)$:

$$\begin{aligned} T(\beta,\theta) &= \hat{T} + B_S(\beta) + B_P(\theta), \\ M(\beta,\theta) &= W(T(\beta,\theta), J(\beta), \theta, \mathcal{W}). \end{aligned} \tag{1}$$

The hand template $T(\beta,\theta) \in R^{778\times3}$, also called the T-template, is obtained by deforming a mean mesh $\hat{T} \in R^{778\times3}$ with shape and pose corrective blend shapes, $B_S(\beta)$ and $B_P(\theta)$. The hand template can be converted into the reference 'rest' T-pose: $J(\beta) \in R^{16\times3}$ with a linear regression using $J(\beta) = \mathcal{J} \times T(\beta,\theta)$, where the $\mathcal{J} \in R^{16\times778}$ matrix stores the regression weights. The blend function $W$ returns the mesh and joint from the T-template, T-pose and $\theta$ parameters, where $\mathcal{W} \in R^{778\times16}$ is a linear blend skinning matrix. With this MANO mesh model, we can easily reconstruct a hand mesh by using specific shape parameters $\beta$ and pose parameters $\theta$.

An often overlooked point for MANO is that the regression weight in the matrix $\mathcal{J}$ is only designed for the rest or T-pose $J(\beta)$. It should not be applied to regress the pose from arbitrary meshes $M(\beta,\theta)$, even though this is done by several existing works [17, 20, 21]. In fact, these works also apply it to non-MANO derived meshes. Our experimental results show a nearly 2 mm gap when $\mathcal{J}$ is used in this way (see Supplementary Section 2).

**Twist-Swing Decomposition:** Directly regressing the pose parameters $\theta \in R^{16\times3}$ is challenging due to the MANO model uses the kinematic chain scheme to adjust joint rotation, that the accumulated error raises the learning challenge [5, 10]. Furthermore, pose parameters $\theta \in R^{16\times3}$ represent the rotation matrix of the hand joints $J$ in SO3 space. However, directly regressing these rotation parameters is an ambiguous problem due to these rotations being non-continuous [41]. The previous work [1] showed that twist-swing decomposition is an effective way for a ball-and-socket joint system to reduce the learning difficulty. Therefore, instead of directly regressing the pose parameters $\theta$, we leverage the twist-swing decomposition and combine the hand joint locations to recover more accurate pose parameters.

### 3.2. Non-parametric Pose Decoder

Like previous works [10, 17, 21], we adopt a ResNet50 [15] as the backbone and encode the input image $I \in R^{256\times256}$ into a latent feature $z \in R^{1000}$. The latent feature $z$ is passed into a pose decoder to obtain the hand joints $J_{\text{pred}} \in R^{21\times3}$.

$$\hat{J} = \mathcal{J} \times M_{\text{np}}(V,F), \quad \text{where} \quad M_{\text{np}}(V,F) = f(z). \tag{2}$$

Specifically, a hand mesh $M_{\text{np}} = (V,F)$ with vertices $V \in R^{778\times3}$ and faces $F \in R^{1538}$ is estimated from the latent feature $z$ with a non-parametric model $f$. The model $f$ can be any off-the-shelf non-parametric method, *e.g.* based on GCNs [10, 17, 21], or transformers [22]. The mesh $M_{\text{np}}$ must have the same topology as the MANO model. Following [17, 20, 21], we use the MANO $\mathcal{J}$ matrix to regress hand joints $\hat{J} \in R^{21\times3}$ from the mesh, even though it is not intended as such and will introduce a performance gap. To bridge this unwanted gap, we add a VAE module to refine the $\hat{J} \in R^{21\times3}$ to $\widetilde{J} \in R^{21\times3}$.

Our proposed VAE module consists of a full connection layer as an encoder and symmetric full connection layers as the decoder. Then, the pose decoder can be learned with the following loss function:

$$L_{\text{joint}} = ||\hat{J} - J|| + ||\widetilde{J} - J|| + ||\hat{J} - \widetilde{J}|| + \lambda_1 * L_{KL} \tag{3}$$

where $J$ is the ground truth joints and $L_{KL} = KL(q(z|\hat{J})||p)$ is the standard Kullback-Leibler divergence loss used in VAE models, where $z$ represents the latent variables encoded from input $\hat{J}$. The term $p = \mathcal{N}(0,E)$ denotes a Gaussian prior where $E$ is an identity matrix.

### 3.3. Parametric Mesh Reconstruction

Directly regressing accurate MANO parameters is highly challenging due to their abstract nature. Twist-swing decomposition is an alternative way to learn these parameters as we already discussed. Further, Li et al. [20] introduced the use of twist-swing decomposition to overcome the challenge of the analogous SMPL [25] model for human body pose and shape estimation. We follow their setting to infer the MANO pose parameters by using the hand joints and

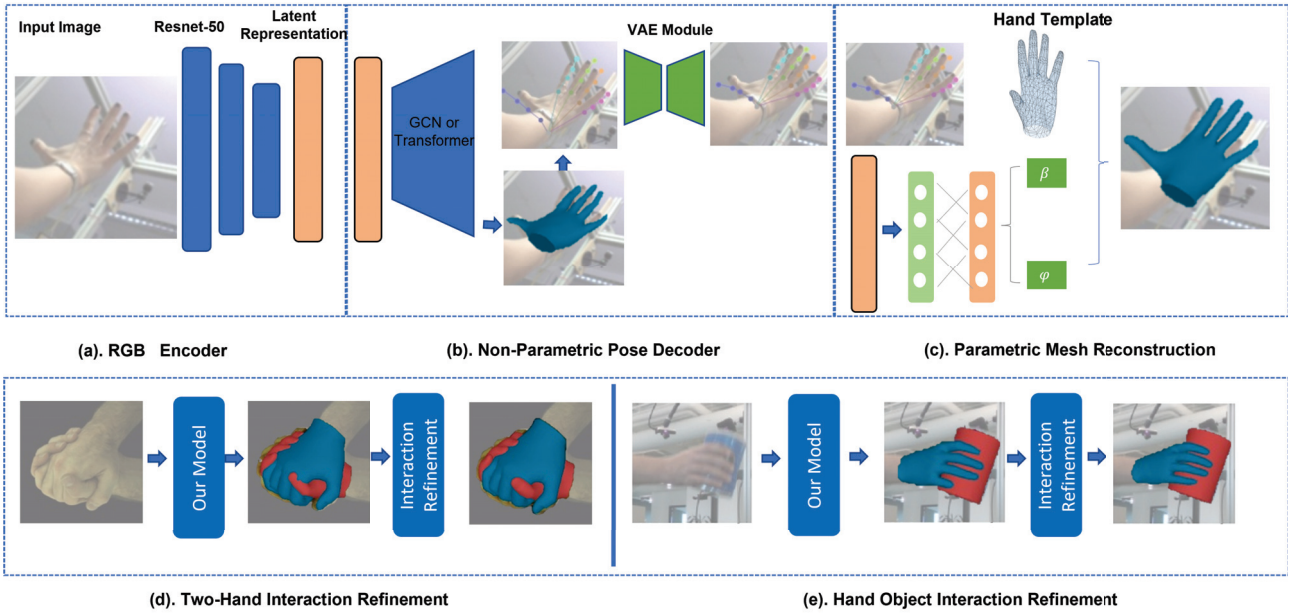


Figure 2. Overview of our pipeline integrating a non-parametric model and a MANO model. Our proposed framework has an RGB image encoder (a), a non-parametric pose decoder (b), and a parametric mesh reconstruction (c). We also show our proposed pipeline success when used in two-hand interaction refinement in (d) and hand-object interaction refinement in (e).

twist parameters, like $\theta = f(\varphi, \widetilde{J})$, where $\varphi \in R^{16\times 3}$ is the hand joint twist rotation matrix.

An interesting finding from our experiments is that there is no difference when using the joint or vertex as supervision for the MANO parameter learning. See Sec. 1 in the Supplementary Material for more details.

These results further demonstrate the possibility of employing joints as weak labels for the learning of MANO mesh. Therefore, we can utilize the estimated joint $J_{\text{pred}}$ and estimated twist rotation matrix $\varphi$ to predict the MANO pose parameter $\theta$. After that, the estimated $\beta$ and $\theta$ are fed into the MANO model to obtain a final well-aligned and plausible hand mesh. For the parametric mesh reconstruction, we use a regularized L1 loss with respect to the ground truth:

$$\begin{aligned} L_{\text{shape}} &= ||\beta_{\text{pred}} - \beta_{\text{gt}}|| + ||\beta_{\text{pred}}||, \\ L_{tw} &= ||\varphi_{\text{pred}} - \varphi_{\text{gt}}|| + ||\varphi_{\text{pred}}||, \end{aligned} \tag{4}$$

where $\beta_{\text{gt}}$ and $\varphi_{\text{gt}}$ are the ground truth shape parameters and twist parameters, respectively.

### 3.4. Interaction Refinement

In our interaction setting, similar to previous works [3, 8, 14, 31], we first estimate the two hands, or the hand and the object, individually, then we add a refinement module. To reduce the penetration for the two-hand interaction, same as [3, 8, 31], we use a Signed Distance Field (SDF) from one hand mesh to check whether the vertex on another hand or object is inside this hand mesh. The SDF is obtained by voxelizing the left and the right hand meshes (two-hand interaction) or hand and object meshes (hand-object interaction) to a $32 \times 32 \times 32$ 3D grid. Then, the modified SDF function $\phi$ for this hand mesh can be written as follows:

$$\phi(x, y, z) = -min(\text{SDF}(c_x, c_y, c_z), 0). \tag{5}$$

For each cell in the 3D grid $c = (c_x, c_y, c_z)$, the $\phi(x, y, z)$ takes positive values if the cell is inside the hand mesh, and zero if outside.

For two-hand interaction, the loss is calculated as follows:

$$L_{\text{pene}} = \frac{1}{|V_{\text{in}}^r|} \sum_{\mathbf{x}_V \in V_{\text{in}}^r} \text{dist}(\text{v}, V_r) + \frac{1}{|V_{\text{in}}^l|} \sum_{\mathbf{x}_V \in V_{\text{in}}^l} \text{dist}(\text{v}, V_l), \tag{6}$$

where $V_{\text{in}}^r$ refers to vertices from the left hand which has penetrated into the right hand, and vice versa for $V_{\text{in}}^l$. The dist$(\cdot)$ represents the minimal distance between the inside vertex and the hand surface. As for the hand-object interaction, we use the same penetration loss in our hand-object interaction refinement following [3],; please refer to [3] for more details.

### 3.5. Weak Label & Self-Distillation

Obtaining ground truth 3D mesh vertices is non-trivial, hence we propose a weakly-supervised approach that uses the 3D joints instead. Recently, self-distillation has become popular for unsupervised pose estimation [19, 24, 29].

We follow a similar approach, under the assumption that the parametric reconstruction, which is strongly supervised with ground truth hand joints and MANO parameters, yields more accurate 3D meshes which can be distilled to the non-parametric branch. We use an L1 loss for self-distillation:

$$L_{\text{vert}} = ||V_{refine} - V_{non}||, \tag{7}$$

where the $V_{refine} \in R^{778\times3}$ and $V_{non} \in R^{778\times3}$ are the hand vertices from our parametric mesh reconstruction module and non-parametric model, respectively. The overall loss function of our proposed pipeline in single-hand shape reconstruction is formulated as:

$$L_{\text{total}} = \lambda_2 L_{\text{shape}} + \lambda_3 L_{tw} + \lambda_4 L_{\text{joint}} + \lambda_5 L_{\text{vert}}. \tag{8}$$

In two-hand interaction and hand-object interaction refinement, we follow [14] and use the above loss function to train the whole pipeline. After that, we use a small learning rate $1e{-}6$ and $L_{\text{pene}}$ to do the interaction refinement; the interaction loss function can be written as follows:

$$L_{\text{inter}} = L_{\text{total}} + \lambda_6 L_{\text{pene}}. \tag{9}$$

# 4. Experimental Results

## 4.1. Implementation Details

Our network consists of three independent modules: image feature extraction, a non-parametric pose decoder and a parametric mesh reconstruction. To ensure a fair comparison, all of our experiments use the same pretrained ResNet 50 [15] as a backbone to extract the input image feature. For the non-parametric pose decoder part, we consider three state-of-the-art structures, *i.e.*, a Graph Convolution Network, a Mesh Transformer Network and a MANO layer network, which are the same as in [6], [22], [14], respectively. The parametric mesh reconstruction module consists of one VAE network and one differentiable layer to calculate the rotation matrix based on given joints. For the shape parameter and twist parameter prediction, two fully-connected networks are used. The above hyper-parameters are set empirically to $\lambda_1 = 0.001, \lambda_2 = 10, \lambda_3 = 10, \lambda_4 = 100, \lambda_5 = 100, \lambda_6 = 10$. The dimension of VAE latent space is 128.

## 4.2. Training Details

The Adam optimizer is applied to train all networks over 200 epochs with a batch size of 64. We start with an initial learning rate of $10^{-4}$ for all training settings and lower it by a factor of 10 at the 50th, 100th, and 150th epochs. After JointVAE is trained, we apply this pretrained JointVAE in our pipeline. We set all of the hypermeters $\lambda$ empirically. In two-hand or hand-object interaction refinement, same as [14], after training the whole pipeline, we use a learning rate of $1e-6$ and a physical contact loss to refine the two-hand and object interaction after the 10th epoch.

## 4.3. Datasets and Evaluation Metrics

**Datasets.** Our method is evaluated on three types of RGB-based hand-object benchmarks, *i.e.*, the One Single hand shape reconstruction dataset on FreiHAND [42], to evaluate the hand shape of our integrated model in single-hand tasks. FreiHAND is a challenging multi-view RGB dataset of hand-object interactions that contains 37k samples of hands manipulating objects. The second one is two-hand shape reconstruction dataset on Interhand2.6M [27] to evaluate our pipeline for mesh reconstruction on a two-hand interaction dataset. Interhand2.6M is a two-hand interaction dataset. We use a dataset setting similar to that of [21], which consists of 366K training samples and 261K testing samples. The last one is hand-object interaction dataset, DexYCB [4], the latest large-scale RGB-based hand-object dataset. It contains 582k samples of hands grasping 20 YCB objects, and is used here to evaluate our proposed model on hand-object interaction. We evaluate our method using the official "S0" split. The hand-object images in this dataset contain 10 objects modeled from YCB objects [34]. We compare our methods with the state-of-the-art on both of these versions and report our results through their leaderboards. All input images are cropped and resized to $256 \times 256$ based on their 2D projection.

**Metrics.** To evaluate the accuracy of our predicted 3D hand pose and surface, we use the mean-per-joint-position-error (MPJPE) for 3D joints and the mean-per-vertex-position-error (MPVPE) for mesh vertices. In addition, unlike previous works that only focus on vertex accuracy for mesh evaluation, we introduce the edge error distance (mm) and normal error [10] as extra evaluation metrics to evaluate the plausibility of hand mesh. Meanwhile, in two-hand interaction, we calculate the penetration depth (PD) of each hand vertices penetrating into the other hand in the 3D grid ($32 \times 32$) above. There are two types of penetration depth, *i.e.*, Average Penetration Depth (A-PD) and Maximum Penetration Depth (M-PD). As for hand-object interaction, the penetration distance calculation is the same as for two-hand interaction.

## 4.4. Comparison with the State-of-the-Art

**Quantitative Results.** The comparison with state-of-the-art non-parametric model-based methods [17, 21, 22] and MANO model based methods [2, 14, 26, 31, 38, 42] is shown in Table 1, where the results is based on their released source code and default parameters. Considering the hand pose and shape accuracy, our integrated model (Ours GCN or Ours Trans) obtains the lowest or second-best MPJPE and MPVPE on all datasets. Especially compared to the latest MANO model based methods, our method reduces the

| Dataset | FreiHAND | | | | InterHand | | | | DexYCB | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Method | MPJPE | MPVPE | Edge | Norm | MPJPE | MPVPE | Edge | Norm | MPJPE | MPVPE | Edge | Norm |
| Zhang et al. [38] | - | - | - | - | 13.48 | 13.95 | 0.31 | 0.15 | - | - | - | - |
| Hasson et al. [14] | 13.3 | 13.3 | 0.68 | 0.17 | 14.21 | - | - | - | - | - | - | - |
| Moon et al. [26] | - | - | - | - | 14.21 | - | - | - | - | - | - | - |
| Rong et al. [31] | - | - | - | - | 17.12 | - | - | - | - | - | - | - |
| Li et al. [21] | - | - | - | - | **8.79** | **9.03** | 0.51 | 0.12 | - | - | - | - |
| Li et al. [21] Baseline | - | - | - | - | 9.97 | 10.63 | 0.51 | 0.12 | - | - | - | - |
| Boukhayma et al. [2] | 13.08 | 13.40 | 0.64 | 0.17 | 16.93 | 17.98 | 0.31 | 0.15 | 12.88 | 12.98 | . 48 | 0.15 |
| MANO CNN [42] | 8.69 | 8.83 | 0.54 | 0.16 | 13.87 | 14.27 | 0.32 | 0.16 | 10.68 | 11.61 | 0.30 | 0.14 |
| GCN-Vert [17] | 7.77 | 7.43 | 0.94 | 0.20 | 9.95 | 10.23 | 0.56 | 0.13 | 8.93 | 9.39 | 0.51 | 0.15 |
| Transformer [22] | 7.57 | 8.05 | 0.81 | 0.16 | 10.89 | 10.83 | 0.68 | 0.15 | 9.51 | 10.48 | 0.65 | 0.15 |
| MANO-Joint [42] | 8.84 | 9.10 | 0.55 | 0.17 | 13.98 | 14.35 | 0.32 | 0.17 | 15.44 | 16.15 | 0.42 | 0.22 |
| GCN-Joint [17] | 14.87 | 18.43 | 3.81 | 0.34 | 11.49 | 19.20 | 5.10 | 0.43 | 10.07 | 15.12 | 3.49 | 0.31 |
| Ours (GCN) | 7.42 | 7.43 | 0.51 | 0.15 | 9.68 | 9.89 | **0.27** | **0.12** | **8.92** | **9.12** | **0.25** | **0.12** |
| Ours (Trans) | **7.28** | **7.33** | **0.49** | **0.15** | 10.08 | 10.06 | 0.29 | 0.13 | 9.13 | 9.67 | 0.28 | 0.14 |

Table 1. Comparisons with state-of-the-art methods on the FreiHAND, InterHand and DexYCB test sets. **Best** and second-best scores. Ours (GCN) and Ours (Trans) achieve the best or second-best holistic performance across all comparisons .

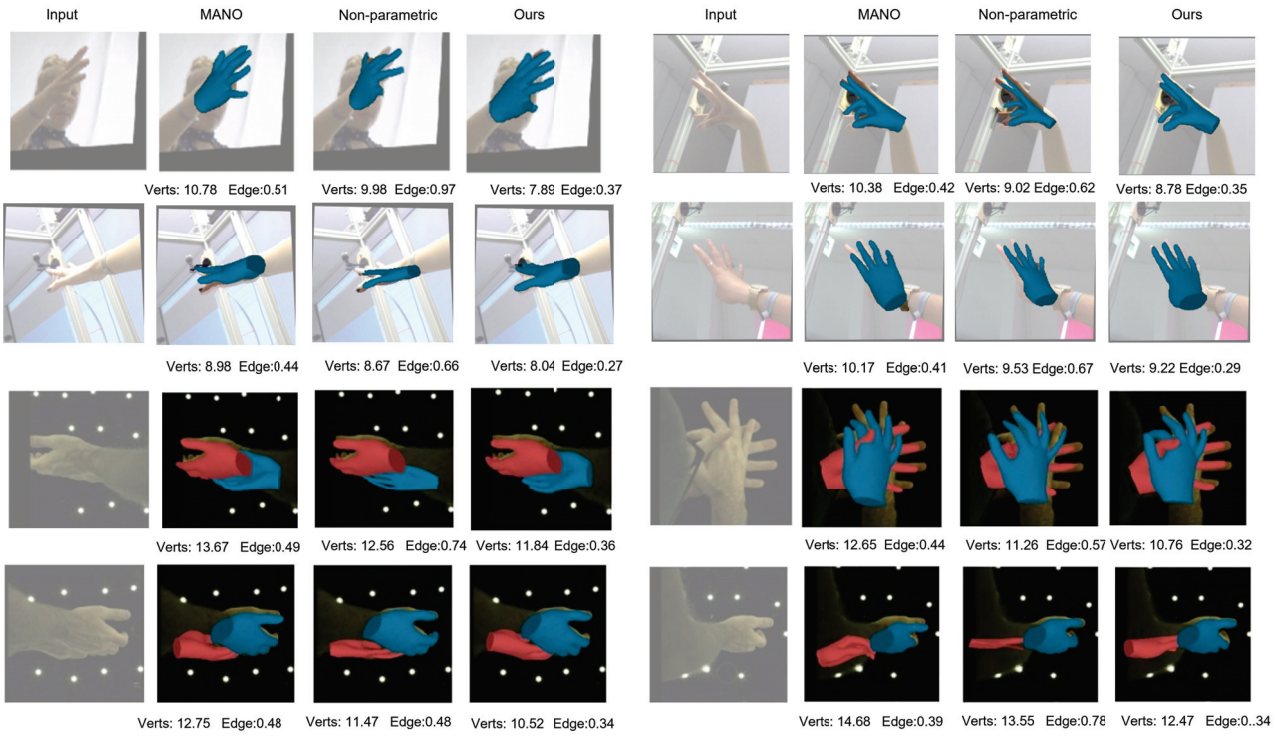


Figure 3. Hand shape reconstruction results. For each quartet, from left to right columns correspond to RGB input, MANO based method: MANO CNN [42], non-parametric model based method: GCN-vert [17] and our method in camera view. Besides, vertex error and edge length error are also reported for quantitative evaluation. Low vertex error and edge length error indicate well-aligned and plausible hand meshes.

pose error MPJPE by nearly 10%. Our MPJPE is comparable to Li et al. [21], despite the attention network used to learn the two-hand features. However, compared to their GCN baseline, our proposed model shows a higher pose and shape accuracy. In addition, regarding the hand mesh plausibility, our proposed model achieves the best holistic

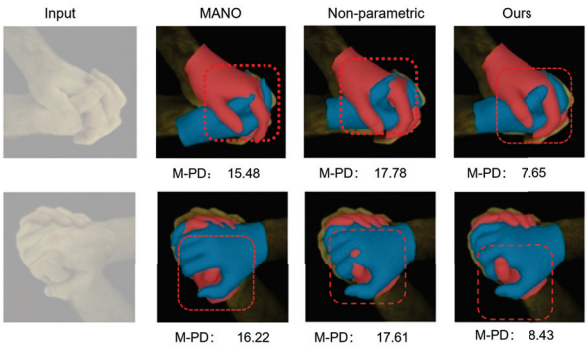


Figure 4. Two-hand reconstruction results. For each quartet, left to right columns correspond to input RGB images, MANO CNN [42], non-parametric model based method [21] and our mesh. The red box highlights the penetration region and we report the max penetration depth values for quantitative evaluation. Our proposed method yields more plausible two-hand interactions.

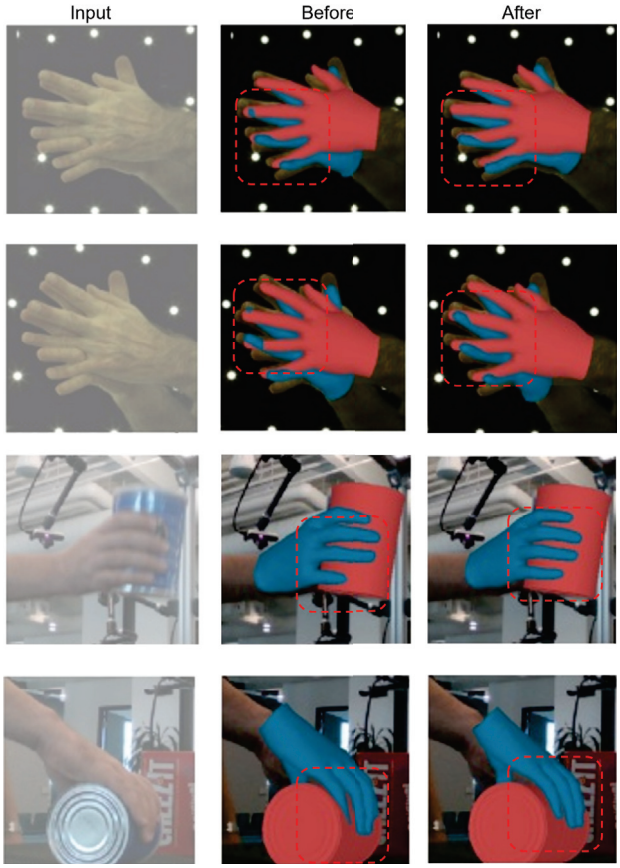


Figure 5. Interaction refinement results. For each triplet, left to right columns correspond to input RGB images, our meshes before and after interaction refinement. Red boxes highlight the interaction refinement regions.

performance in terms of edge distance and normal error across all comparisons. Especially compared to the non-parametric model based methods, our method reduces the edge distance error by at least 40%. The above quantitative results verify the effectiveness of our integrated model in obtaining well-aligned and plausible hand meshes.

**Qualitative Results.** The visualizations of our hand modeling in Fig. 3 verify that our proposed model achieves well-aligned and plausible hand reconstructions. Additionally, Fig. 4 compares our method to state-of-the-art, demonstrating that our proposed method has lower penetration and yields more plausible two-hand interactions, although [21] achieved better MPJPE and MPVEP than our method by using a complex attention network. Besides, these visualization results also verify that the physical contact loss refinement is better than feature level refinement by using an attention network. More qualitative results are available in the Supplementary.

**Interaction Refinement Results.** We also show the quantitative results in Table 2 and qualitative results in Fig 4. Our model (Ours Before) achieves the best or second-best performance across all comparisons. In addition, as our proposed model integrates the MANO model, we can leverage the physical contact loss (Ours After) to refine our two-hand or hand-object interactions and reduce M-PD by nearly 50%. This reveals the effectiveness of our proposed interaction refinement and emphasizes the importance of physical contact loss when compared to the attention feature learning like [21] for interaction refinement.

| Dataset | InterHand | | | DexYCB | | |
|---|---|---|---|---|---|---|
| Method | MPVPE | A-PD | M-PD | MPVPE | A-PD | M-PD |
| Li et al. [21] | 9.03 | 1.04 | 17.61 | - | - | - |
| MANO-CNN [42] | 14.27 | 1.03 | 17.81 | 11.61 | 1.01 | 16.78 |
| GCN-Vert [17] | 10.23 | 1.03 | 17.95 | 9.39 | 0.98 | 16.95 |
| Transformer [22] | 10.83 | 1.04 | 18.37 | 10.48 | 1.05 | 17.54 |
| GT | 0 | **0.17** | **4.89** | 0 | **0.15** | **3.21** |
| Ours (Before) | 9.89 | 1.00 | 17.51 | 9.12 | 0.94 | 16.51 |
| Ours (After) | **9.92** | 0.51 | 7.62 | **9.33** | 0.45 | 6.73 |

Table 2. Comparisons between our model (before and after refinement) versus state-of-the-art on InterHand2.6M and DexYCB test sets. **Best** and second-best scores. Our model achieves the best interaction performance across all comparisons. Note that the A-PD and M-PD of InterHand and DexYCB ground truth data are non-zero due to the rigid modeling of both the hand and the object.

### 4.5. Ablation Studies

**VAE Module.** We also compare among our baseline models (Ours w/o VAE) in Table 3. Our model (w/o VAE) is a pipeline without refining the joints, which directly use a linear regress matrix to covert non-parametric model

| Dataset | FreiHAND | | | |
|---|---|---|---|---|
| Method | MPJPE | MPVPE | Edge | Norm |
| Ours(w/o VAE) | 9.13 | 9.20 | 0.60 | 0.18 |
| Ours(w/o Self-dis.) | 8.63 | 8.66 | 0.58 | 0.16 |
| Ours (full) | **7.42** | **7.43** | **0.51** | **0.15** |

Table 3. Ablation study on FreiHAND test sets. Best scores are highlighted in **Bold**.

meshes to the MANO model joints space. Our full model outperforms this baseline by nearly 20%, which verifies the effectiveness of the VAE refinement module.

**Self-distillation learning.** The impact of self-distillation learning is shown in Table 3 (Ours w/o Self-dis.). Self-distillation learning reduces pose and shape error by nearly 10%. Furthermore, our full model reduces the pose and shape errors by nearly 50% compared to the non-parametric model based method, *i.e.*, GCN-joint in Table 1, which only uses joint as supervision. This reveals the effectiveness of our self-distillation learning and integrated strategy.

**Analysis of the twist rotation.** To evaluate the effectiveness of the twist rotation, besides our estimated twist from the network (Estimated Twist), we also set the twist as zero (Zero Twist) or a random value from 0 to 1 (Random Twist). The results are given in Table 4. Firstly, the performance of the Zero Twist is comparable to that of our Estimated Twist. These results are reasonable since most hand joints' twist rotation angles are close to zero. In contrast, there is a considerable performance gap between the Random Twist results and our Estimated Twist results, which shows the necessity of twist rotation estimation.

| Dataset | FreiHand | | DexYCB | |
|---|---|---|---|---|
| Method | MPJPE | MPVPE | MPJPE | MPJPE |
| Random Twist | 12.68 | 13.90 | 11.78 | 12.52 |
| Zero Twist | 7.83 | 7.96 | 9.45 | 9.67 |
| Estimated Twist | **7.42** | **7.43** | **8.92** | **9.12** |

Table 4. Reconstruction error with different twist angles. Best scores are highlighted in **Bold**.

### 4.6. Limitations

Our pose decoder pipeline uses a non-parametric model to obtain the initial meshes. These initial hand meshes limit our results compared to the ground truth (see Fig. 6). Although our initial hand meshes are not well aligned, our output meshes are close to the ground truth and better than our initial hand meshes, which verifies the effectiveness of our integrated model on the other side. These limitations can be improved by considering other sources of information, like rendered masks, to offer extra supervision to improve the initial hand mesh accuracy.

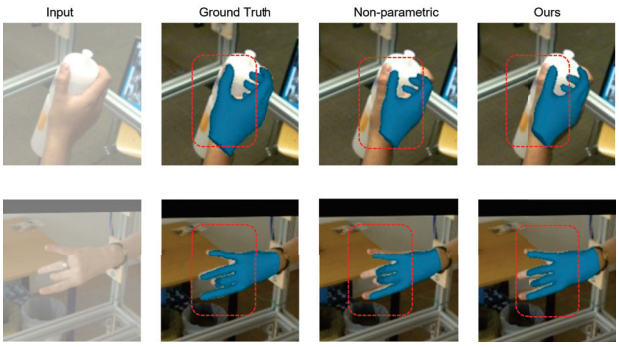


Figure 6. Limitation results. For each row, the left to right columns correspond to input RGB, ground truth, our non-parametric model mesh and our final mesh. We are limited by our initial hand mesh from our non-parametric model. Red boxes highlight the not aligned regions.

## 5. Conclusion

This work proposes an effective integrated framework of a non-parametric model and MANO model for estimating well-aligned and plausible hand meshes from RGB images. We explore the trade-off between the non-parametric and MANO model for hand surface modelling and propose the first integrated model to overcome this trade-off. Additionally, to improve the accuracy of hand meshes and mitigate the gap between the non-parametric model joints and MANO deformation joints, we introduce a VAE to solve it. Furthermore, we introduce a self-distillation learning method that utilizes our parametric mesh to boost the non-parametric model's mesh learning. Experimental results show that our proposed method achieves better performance over existing MANO-based and non-parametric model based hand shape estimation methods, on single-hand task, two-hand interaction or hand-object interaction task. This verifies the effectiveness of our integrated framework of a non-parametric model and a parametric model. In future work, we will explore using a render mask as extra supervision to improve the hand shape modeling based on our integrated model.

**Acknowledgements** This research is supported by the National Research Foundation, Singapore and DSO National Laboratories under its AI Singapore Programme (AISG Award No: AISG2-RP-2020-016). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

## References

[1] Paolo Baerlocher and Ronan Boulic. Parametrization and range of motion of the ball-and-socket joint. In *Deformable*

*avatars*. 2001. 2, 3

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 1, 2, 5, 6

[3] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 1, 2, 4

[4] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 5

[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 *ICCV*, 2021. 2, 3

[6] Xingyu Chen, Yufeng Liu, Dong Yajiao, Xiong Zhang, Chongyang Ma, Yanmin Xiong, Yuan Zhang, and Xiaoyan Guo. Mobrecon: Mobile-friendly hand mesh reconstruction from monocular image. In *CVPR*, 2022. 1, 2, 3, 5

[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 *CVPR*, 2021. 1, 2

[8] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. *ECCV*, 2022. 1, 2, 4

[9] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *NeurIPS*, 2016. 2

[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 *CVPR*, 2019. 1, 2, 3, 5

[11] Shunwang Gong, Lei Chen, Michael Bronstein, and Stefanos Zafeiriou. Spiralnet++: A fast and highly efficient mesh convolution operator. In *ICCV*, 2019. 1, 2

[12] 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*, 2022. 2

[13] 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 *CVPR*, 2020. 2

[14] Yana Hasson, Gül Varol, Dimitris Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 4, 5, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2015. 3, 5

[16] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2

[17] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 1, 2, 3, 5, 6, 7

[18] Chen Li and Gim Hee Lee. Coarse-to-fine animal pose and shape estimation. In *NeurIPS*, 2021. 1

[19] Chen Li and Gim Hee Lee. From synthetic to real: Unsupervised domain adaptation for animal pose estimation. In *CVPR*, 2021. 4

[20] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *CVPR*, 2021. 1, 2, 3

[21] 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*, 2022. 1, 2, 3, 5, 6, 7

[22] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 1, 2, 3, 5, 6, 7

[23] Shuying Liu, Wenbin Wu, Jiaxian Wu, and Yue Lin. Spatial-temporal parallel transformer for arm-hand dynamic estimation. In *CVPR*, 2022. 2

[24] Zhidan Liu, Zhen Xing, Xiangdong Zhou, Yijiang Chen, and Guichun Zhou. 3d-augmented contrastive knowledge distillation for image-based object pose estimation. *arXiv preprint arXiv:2206.02531*, 2022. 4

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *SIGGRAPH Asia*, 2015. 2, 3

[26] 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*, 2020. 5, 6

[27] 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*, 2020. 1, 2, 5

[28] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, 2022. 1, 2

[29] Pengfei Ren, Haifeng Sun, Weiting Huang, Jiachang Hao, Daixuan Cheng, Qi Qi, Jingyu Wang, and Jianxin Liao. Spatial-aware stacked regression network for real-time 3d hand pose estimation. *Neurocomputing*, 2021. 4

[30] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH Asia*, 2017. 1, 2, 3

[31] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 4, 5, 6

[32] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*, 2022. 2, 3

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2

[34] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 5

[35] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 2

[36] Ziwei Yu, Linlin Yang, Shicheng Chen, and Angela Yao. Local and global point cloud reconstruction for 3d hand pose estimation. In *BMVC*, 2021. 2

[37] Ziwei Yu, Linlin Yang, You Xie, Ping Cheng, and Angela Yao. Uv-based 3d hand-object reconstruction with grasp optimization. In *BMVC*, 2022. 2

[38] 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 *ICCV*, 2021. 1, 2, 5, 6

[39] Jason Y. Zhang, Sam Pepose, Hanbyul Joo, Deva Ramanan, Jitendra Malik, and Angjoo Kanazawa. Perceiving 3d human-object spatial arrangements from a single image in the wild. In *ECCV*, 2020. 2

[40] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *CVPR*, 2019. 1, 2

[41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and ao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2018. 3

[42] 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*, 2019. 1, 5, 6, 7