Interacting Two-Hand 3D Pose and Shape Reconstruction from Single Color Image

Baowen Zhang[1,2] Yangang Wang[3] Xiaoming Deng[1,2*] Yinda Zhang[4*]
Ping Tan[5,6] Cuixia Ma[1,2] Hongan Wang[1,2]
[1]Institute of Software, Chinese Academy of Sciences [2]University of Chinese Academy of Sciences
[3]Southeast University [4]Google [5]Simon Fraser University [6]Alibaba

Abstract

In this paper, we propose a novel deep learning framework to reconstruct 3D hand poses and shapes of two interacting hands from a single color image. Previous methods designed for single hand cannot be easily applied for the two hand scenario because of the heavy inter-hand occlusion and larger solution space. In order to address the occlusion and similar appearance between hands that may confuse the network, we design a hand pose-aware attention module to extract features associated to each individual hand respectively. We then leverage the two hand context presented in interaction to propose a context-aware cascaded refinement that improves the hand pose and shape accuracy of each hand conditioned on the context between interacting hands. Extensive experiments on the main benchmark datasets demonstrate that our method predicts accurate 3D hand pose and shape from single color image, and achieves the state-of-the-art performance. Code is available in project webpage https://baowenz.github.io/Intershape/.

1. Introduction

3D hand pose and shape reconstruction plays an important role in many applications, such as AR/VR [8] and robotics [9]. While most of the previous hand pose and shape reconstruction works [3, 41] are proposed for single hand, we study the problem of hand reconstruction for two interacting hand from single color image, as it is more desirable to express delicate body language [39] and perform complex tasks [18, 25, 36]. However, the prior art on this topic is barely missing. Existing methods usually rely on depth sensor [19], multi-view camera system [8] or optimization over tracked motion sequence [19, 8], which however are either relative expensive, energy consuming, or sensitive to the tracking quality and initialization. Comparatively, single color camera setup is more cost and computation friendly, and it is also widely available. Therefore, we focus on conducting interacting two-hand reconstruction from single color image (See Fig. 1).

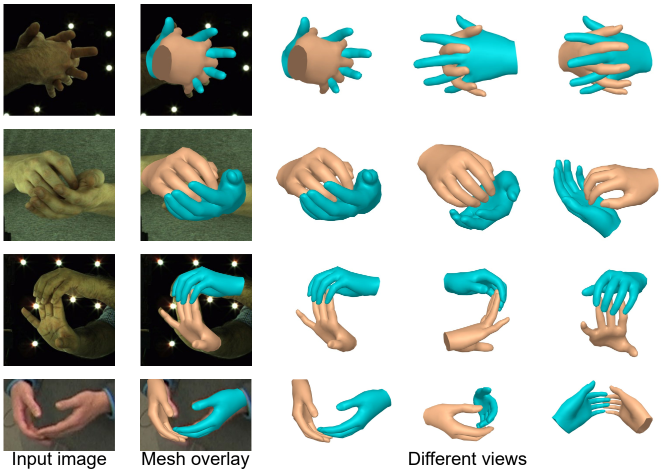


Figure 1. Illustration of interacting hand shape reconstruction from single color image. Our method can get high-quality reconstruction under heavy interhand occlusions.

Reusing similar hand reconstruction techniques designed for single hand for the two-hand scenario is non-trivial. First, compared to the case with a single complete hand, two hands are usually heavily occluded and tightly contacting with each other due to the interaction, which are much harder to parse. Two hands also share similar textures, which can easily confuse the network to extract feature from correct regions in the image. Second, the ill-pose nature of the problem is exacerbated with the degree of the freedom of the solution space doubled. The model is error-prone and may produce two hands in unreasonable pose and shape that people would rarely or be infeasible to present.

Recently, Moon *et al.* [18] propose a large-scale interacting hand dataset named InterHand2.6M, and present an interacting hand pose estimation method. However, less

[*]indicates corresponding author

special design is conducted to handle the characteristic of two-hand pose estimation problem, and more fine-grained hand shape reconstruction is also not explored in [18].

To address the above mentioned issues, we propose a novel deep learning architecture for interacting hand pose and shape estimation (See Fig. 2). Our network consists of an encoder that extracts multi-scale features, and a decoder to gradually refine the prediction with feature at each level. In the encoder, a per-hand heatmap is estimated and used to mask the image features, which is particularly effective to extract the features from correct image regions and produce accurate prediction for each individual hand. On the other hand, the decoder is designed to leverage the context between interacting hands. Instead of optimizing each hand separately, we refine each hand conditioned on the current estimation two hands. Our network generates attention map to reduce feature ambiguity between two hands. Different from traditional methods that generate attention map from features in a network, we propose to generate attention map directly from estimated hand shape. In order to jointly recover hand skeleton pose and shape, we adopt the popular hand statistical model MANO [27] and predict the MANO parameters of two hands respectively.

Our main contributions are summarized as follows:

1. We propose a novel deep learning architecture, which can estimate 3D hand pose as well as fine-grained hand shapes of the interacting hands from single color image. Our work can also inspire several related researches such multiple person reconstruction, hand-object interaction reconstruction etc;

2. In order to address the feature ambiguity between two hands, we propose pose-aware attention modules to extract the key features for each hand;

3. We leverage the two-hand context presented in interaction and propose a cascaded refinement stage improving the hand pose and shape accuracy of each hand conditioned on context of interacting hands;

4. Extensive experiments shows that our method achieves state-of-the-art performance on the main datasets.

2. Related Work

Single Hand Pose and Shape Reconstruction from Color Due to the advantage of ubiquity and low power consumption of color image, it is highly desired to recover 3D hand pose from color images. Prior art works on 3D hand pose estimation include [42, 31, 11, 38, 30, 4, 5]. Most of the hand pose and shape reconstruction methods from color use a parametric model such as MANO [27] to represent hand shape, and learn the hand shape model parameters from image. Boukhayma *et al*. [3] use 2D pose, 3D pose and hand mask as supervisions to train the hand shape network. Zhang *et al*. [40] also design a hand reconstruction network to learn MANO parameters using the predicts 2D heatmap and the image feature. Zhou *et al*. [41] estimate MANO shape parameter from predicted 3D pose. Moon *et al*. [17] propose a weakly-supervised model to reconstruct hand shape, which does not require any ground truth hand meshes. Different to these methods using MANO model, Ge *et al*. [7] present a graph neural network for hand shape reconstruction, which can capture local geometry details well. Han *et al*. [8] propose a tracking-based approach to estimate 3D hand poses using four fisheye monochrome cameras. In order to address the lack of large-scale hand reconstruction dataset, Zimmermann *et al*. [43] present a multi-view single hand dataset with both 3D hand pose and shape annotations. Kulon *et al*. [14] annotate a hand shape dataset using model fitting. However, these hand reconstruction methods are proposed for single hand, and they does not explicitly address overlapping or interacting hands.

Interacting Hand and Object Shape Reconstruction Hand shape reconstruction is related to the hand-object shape reconstruction. Prior art works include [10, 9, 21, 6, 1]. Compared to hand-object shape reconstruction, interacting two-hand reconstruction is more difficult, because it aims to reconstruct two interacting articulated hands, which leads to more inter-occlusions, deformations, and motions of degree-of-freedom. Moreover, for hand-object interaction, the hand and object have additional contact constraints to model the hand-object relationship. However, for interacting hand reconstruction, two hands may not have contacts and result in larger solution space.

Interacting Hand Pose and Shape Estimation Most of existing works conduct two-hand reconstruction by multiple cameras [2, 8], a single depth camera [19, 22, 33, 15, 32] and tracking strategy [22, 15, 35, 29]. Due to the ubiquitous characteristic of single color image, the methods using single color image are more preferable than tracking methods, methods using multi-cameras and depth camera. Moon *et al*. [18] propose the InterHand2.6M dataset for single and interacting hand pose estimation, and use the dataset to train a network to predict 2.5D hand poses for two hands. Lin *et al*. [16] use a synthetic dataset to learn two-hand pose from single color image. However, these methods can not achieve satisfactory two-hand pose estimation results or reconstruct fine-grained geometry – hand shapes.

Full Body Reconstruction Full body reconstruction methods [39, 13, 24, 28] implicitly handle two-hand reconstruction. However, these methods require that most of body parts are visible. Existing full body methods do not

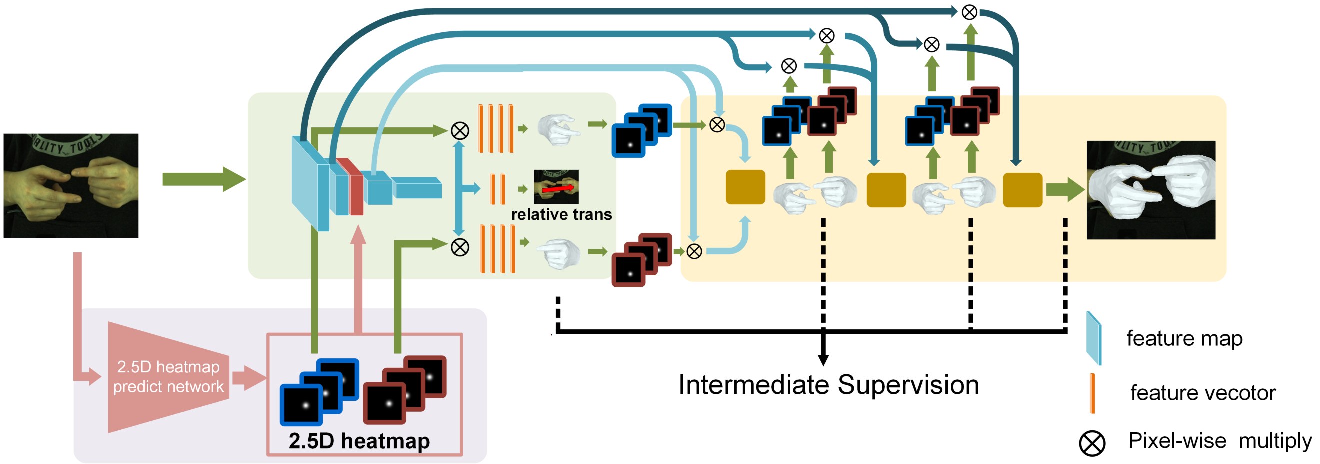


Figure 2. Illustration of our interacting hand shape reconstruction network. Our network first predict 2.5D heatmap for the joints of the two hands. Then use three branches to recover MANO model parameters of each hand and the relative transformation of two hands. Finally, refine the hand shape parameters jointly in a cascaded manner to respect the correlation context between the interacting hands.

contain special modules to handle the distinct characteristics of close hand interaction. Interacting hand reconstruction is more challenging than full body reconstruction, because less context of body parts is available to reduce the reconstruction ambiguity due to interhand occlusions.

Different to these methods, our method presents a novel deep learning method to predict interacting two-hand pose and shape from single color image. We adopt a pose-aware attention module to help the network to learn the features relevant to each hand. In order to resolve the ambiguity due to the interhand occlusions, we leverage the context between interacting hands to refine the pose and shape of two hands using a cascaded network.

3. Method

In this section, we introduce our model for interacting two-hand pose and shape reconstruction from a single color image. An overview of our model is shown in Fig. 2. Our model starts from a multi-scale feature extractor built upon ResNet-50 architecture. Inspired by InterHand [18], we predict a per-joint heatmap for both hands, and inject it to the encoder for feature extraction. The feature in the lowest resolution is then fed into a network to produce an initial estimation for the shape and pose of two hands and their relative transformation. The shape and pose of two hands are then jointly refined in a cascaded manner by leveraging features in high resolutions. This refinement stage learns context between interacting hands and is effective to improve the hand reconstruction quality.

3.1. Interacting Hand Representation

We use the statistical hand model MANO [27] to represent the hand shape and hand pose of two hands. The hand surface mesh $\mathbf{M}$ can be deformed with hand shape parameter $\beta \in R^{10}$ and hand pose parameter $\theta \in R^{16\times3}$

$$\mathbf{M} = W(\mathbf{T}(\beta,\theta), \mathbf{J}(\beta), \mathbf{W}) \tag{1}$$

where $W(\cdot)$ is skinning function, $\mathbf{T}$ is a parametric hand template shape, $\mathbf{J}(\beta)$ is the hand joint position at the rest pose, and $\mathbf{W}$ is a skinning weight matrix.

For our two-hand interacting scenario, the goal is to predict MANO parameters including the pose parameters and shape parameters for both hands, i.e. $(\beta_{left}, \theta_{left})$ and $(\beta_{right}, \theta_{right})$, and the relative translation Δ and the scale s between two hands. The output MANO models of our method are aligned with the root joint of each individual hand, and the root rotations of both hands are in the camera coordinate system. We use the output relative translation Δ and the scale s of our model to merge the hand pose results from hand reconstruction network of two hands as follows

$$\mathbf{J}^{right}_{left,i} = s(\mathbf{J}^{left}_{left,i} + \Delta) \tag{2}$$

where $\mathbf{J}^{right}_{left,i}$ and $\mathbf{J}^{left}_{left,i}$ are the left hand joints in the right hand coordinate system and the left hand joints in the left hand coordinate system, respectively. The hand shape reconstruction results can be also merged in a similar way.

3.2. Pose-aware Feature Extractor

Particularly for interacting hand scenario, it is important to provide features for each individual hand to ensure accurate reconstruction respectively. Traditional methods generate attention map from features in network. 2D/2.5D heatmap estimation network, e.g stacked hourglass [20] and SRNet [37], can directly obtain 2D attention map. However, these methods cannot be used in 3D shape estimation

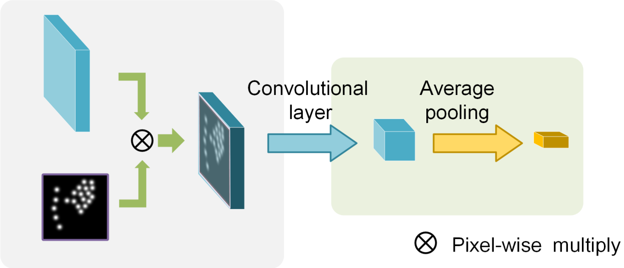


Figure 3. Illustration of our feature extraction module using attention map. The feature map is multiplied with attention map and is down-sampled using convolutional layer and average pooling.

scenario. To this end, we perform a pose-aware feature extractor using an attention map, which identifies region of interests for each hand. The attention map is multiplied with each channel of the feature maps, and is down-sampled via multiple convolutional layers. The low-resolution feature is fed into a global average pooling to extract per-hand feature vector, which will be used for hand reconstruction. Fig. 3 illustrates the feature extraction process.

Instead of learning in a black-box, we resort to per-joint heatmap to produce the attention map for each individual hand. The value on each pixel of the attention map measures the probability of the existence of any hand joints:

$$\mathbf{A}_h = 1 - \prod_{i=1}^{K}(1 - \mathcal{H}_{h_i}), \quad h \in \{right, left\}, \qquad (3)$$

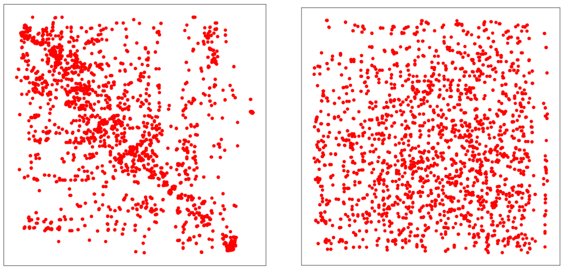
where $\mathcal{H}_{h_i}$ is the heatmap for joint i in hand h, and K is the total number of hand joints.

3.3. Interacting Hand Reconstruction

We then introduce how to predict pose and shape for interacting hands using extracted feature maps from the encoder. Our method starts from an initial estimation followed by a context-aware model to jointly refine the results.

Initial Estimation To predict an initial estimation of the interacting hands, we take the feature map in the lowest resolution, and extract pose-aware feature vectors for each hand. In order to get the attention map, we use the predicted 2.5D heatmap as [18], concatenate the 2.5D heatmap of each joint in depth dimension, and conduct max-pooling along the channel dimension. We also generate a feature for the relative transformation by directly applying an average pooling since it requires information from both hands. The feature vectors are then fed into separate MLP to predict the MANO parameter for each hand respectively, and the relative translation and scale.

Context-Aware Refinement We then refine the initial estimation with high-resolution features provided by the encoder, which contains more spatial information that may benefit the network to recover details. While the pose-aware feature extractor resolves the two-hand ambiguity and provides more specific feature for each hand, it loses the chance to jointly optimize two hands leveraging the context.



(a) Paired hand poses (b) Unpaired hand poses

Figure 4. Visual analysis of 2D manifold of paired (a) and unpaired (b) two hand poses. The paired hand poses show clear correlation in 2D space, but the distribution of unpaired hand poses is almost random. The paired hand poses are sampled from InterHand2.6M [18], and the unpaired poses are permuted from the paired ones.

In fact, left and right hands show strong correlation when interacting with each other. To show this, we conduct visual analysis of paired and unpaired two hand poses (Fig. 4). The paired hand poses are sampled from InterHand2.6M [18], and the unpaired poses are permuted from the paired ones. Inspired by [26], we use 2D manifold representation, where the hand pose (without root rotation) of each hand is projected to 1D manifold by t-SNE [34] and used as x, y coordinates, respectively. We find that the paired hand poses show clear correlation in 2D space, but the distribution of unpaired hand poses is almost random.

Inspired by this, we design a cascaded refinement stage that jointly optimizes two hands. For a particular hand, we first render the hand joints heatmaps according to the estimated MANO parameters in the previous stage. Each joint is projected onto the input image and rendered as a 2D Gaussian map with variance as 1.5. In order to render the hand joint heatmap, we obtain the weak perspective camera parameters by aligning the 3D joint positions from the predicted MANO parameters and the predicted 2.5D heatmap. We then extract the pose-aware feature using these heatmaps, concatenate it with the MANO parameters estimated for both hands in the last stage, and feed them into an MLP to produce updated MANO parameters. To gradually bring in detailed spatial information, the later refinement stage uses features in higher resolution coming from early layers in the encoder.

3.4. Loss Functions

We add loss functions to supervise intermediate and final network outputs. Specifically, we add losses for attention map, and the MANO parameters for left and right hand before and after context-aware refinement.

3.4.1 Two Hand Loss

Joint Offset Loss In order to enforce the relative position of the corresponding joints of both hands, we supervise the offsets of the corresponding joints of two hands

$$L_o = \sum_{i=1}^{K} ||(\mathbf{J}_{right\ ,i} - \mathbf{J}_{left\ ,i}) - (\mathbf{J}^*_{right\ ,i} - \mathbf{J}^*_{left\ ,i})||_2^2 \tag{4}$$

where $\mathbf{J}_{right\ ,i}$ and $\mathbf{J}_{left\ ,i}$ are the estimated i-th joints of two hands, $\mathbf{J}^*_{right\ ,i}$ and $\mathbf{J}^*_{left\ ,i}$ are the ground truth.

Shape Consistence Loss Due to the symmetry of two hands of a subject, the hand shape parameters of two hands should be close. Thus, we use L2 distance of $\beta_{left}, \beta_{right}$ to enforce the hand shape consistency of two hands

$$L_c = ||\beta_{right} - \beta_{left}||_2^2 \tag{5}$$

3.4.2 Single Hand Loss

Joint Loss We use L1 distance between the ground truth hand joints and the predicted hand joints

$$L_J = \sum_{h \in \{left, right\}} \sum_{i=1}^{K} ||\mathbf{J}_{h,i} - \mathbf{J}^*_{h,i}||_1 \tag{6}$$

where $\mathbf{J}_{h,i}$ and $\mathbf{J}^*_{h,i}$ are the predicted and ground truth positions of the i-th joint, and K is the number of hand joints.

Bone Length Loss We use L2 loss to supervise the predicted bone length. Since we predict scale-normalized hand pose and bone length (relative to the length of a reference bone l_{ref} connecting the MCP joint of the middle finger and the wrist joint), we use the Euclidean distance between the normalized ground truth bone lengths and the predicted bone lengths to calculate the bone length loss

$$L_l = \sum_{h \in \{left, right\}} \sum_{b} ||\frac{l^*_{h,b}}{l^*_{h,ref}} - l_{h,b}||_2^2 \tag{7}$$

where $l^*_{h,b}$ and $l^*_{h,ref}$ are the ground truth bone lengths for the b-th bone and the reference bone, respectively.

Shape Loss We use shape loss to enforce the hand shape parameters

$$L_M = \sum_{h \in \{left, right\}} \mathbf{1}||\beta_{\mathbf{h}} - \beta^*_{\mathbf{h}}||^{\mathbf{2}}_{\mathbf{2}} \tag{8}$$

which aims to enforce the predicted MANO shape parameters close to the ground truth. $\mathbf{1}$ is an indicator function that is 1 if ground truth MANO shape parameters are labeled and 0 otherwise, and β^*_{right} and β^*_{left} are the ground truth of the hand shape parameters.

Regularizer Loss We use regular terms to enforce predicted MANO parameters remaining reasonable

$$L_{reg} = \sum_{h \in \{left, right\}} \lambda_\beta ||\beta_h||_2^2 + ||\theta_h||_2^2 \tag{9}$$

where the shape regularizer $||\beta_h||_2^2$ enforces the reconstructed shapes to be close to the mean shape (i.e. β=0), and the pose regularizer $||\theta_h||_2^2$ [27] helps eliminate joint twist. The loss weight λ_β is set to 0.1.

The total loss function of our network is defined as follows:

$$\begin{aligned} L_{total} = &\lambda_o L_o + \lambda_c L_c + \lambda_J L_J \\ &+ \lambda_l L_l + \lambda_M L_M + \lambda_{reg} L_{reg} \end{aligned} \tag{10}$$

where $\lambda_o, \lambda_c, \lambda_J, \lambda_l, \lambda_M, \lambda_{reg}$ are the loss weights, and they are set to 1, 0.01, 10, 100, 0.1, and 0.05, respectively.

3.5. Implementation Details

We implement our network with Pytorch [23]. We use Adam optimizer to train our network. The learning rate is set to 5×10^{-5}, the mini-batch size is set to 20, and the training iteration is set to 500K.

We follow prior art [18] to crop hand region with the annotated bounding box in the training and testing dataset. The images are resized to 256 × 256. To achieve scale-invariant shape estimation, we normalize the distance from the middle finger MCP joint to the wrist joint to 1. During the testing stage, we used the ground truth bone lengths of the two hands to recover their scales. During the training stage, we do not give direct supervisions to Δ and s, but use offset loss and joint loss to enforce them (See Sec. 3.4).

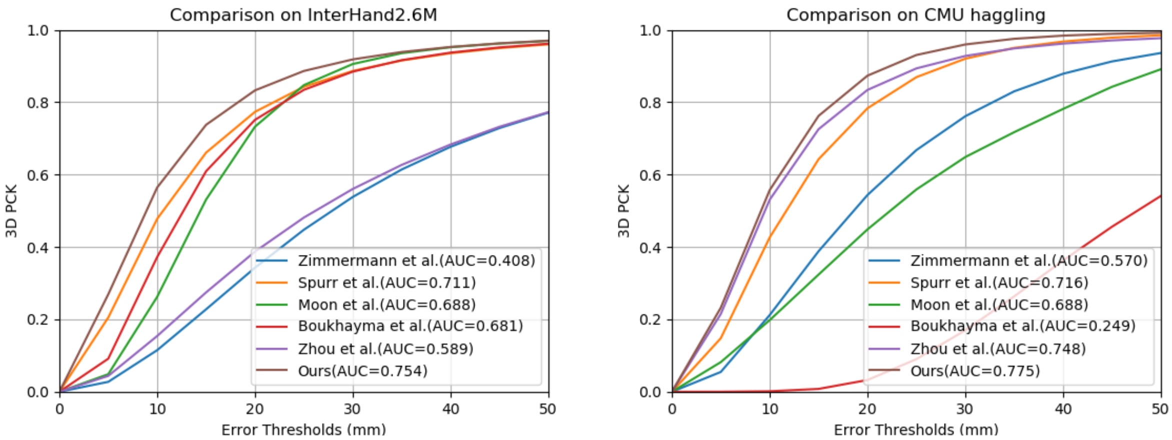
4. Experiment

We conduct experiments on two main benchmark datasets to verify the performance of our method. We compare our method to the state-of-the-art methods (See Sec. 4.2), and adopt the ablation study to evaluate the effect of each component of our method (See Sec. 4.3).

4.1. Datasets and Evaluation Metrics

We evaluate the performance of our method on the popular two-hand reconstruction benchmark datasets, InterHand2.6M [18] and Haggling [12]. Other recently proposed two-hand datasets with relatively less high-quality data, such as RGB2Hands (a dataset containing unrealistic data without background) [35] and Ego3DHands (a synthetic dataset with domain gap to real data) [16] do not satisfy the requirements of hand reconstruction evaluation.

InterHand2.6M [18] consists of 2.6M labeled single and interacting hand frames under different poses from multiple subjects. We adopt the interacting hand frames (IH) in the



(a) Comparison on InterHand2.6M dataset

(b) Comparison on Haggling dataset

Figure 5. Quantitative comparison of our method to the state-of-the-art methods.

dataset for training and evaluation. The training and testing dataset contain 141,497 and 125,689 frames, respectively.

Haggling Dataset [12] This dataset contains multiple sets of videos of haggling games. People's gestures in the videos contain a lot of semantic information. We keep two-hand interacting frames that are accurately labeled as training and testing data. The dataset is divided into training set and testing set according to [12]. The training and testing dataset contain 80,953 and 24,363 frames, respectively.

Evaluation Metrics To evaluate the accuracy of 3D hand pose estimation, we follow the prior art [18] to use the mean per joint position error (MPJPE) in millimeters, and also adopt the Percentage of Correct Keypoints (PCK), both after root joint alignment. The root joint alignment is performed for left and right hand separately as [18]. We also evaluate the hand pose performance using Area Under the Curve (AUC) (0-50) mm of PCK curve over different error thresholds. To evaluate the accuracy of 3D shape reconstruction, we use the mean error between the corresponding vertices of ground truth and predicted hand shapes as evaluation metric, named shape-err, after root joint alignment of each hand. In the qualitative results, we align the root joint of each hand with the ground truth in camera coordinate for visualization.

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

Firstly, we compare the hand pose performance to the state-of-the-art two-hand pose estimation method InterHand [18], and single hand reconstruction methods, including Zimmermann *et al.* [42], Spurr *et al.* [31], Boukhayma *et al.* [3], and Zhou *et al.* [41]. For the two-hand approach InterHand [18], we directly use their output 3D pose for evaluation on InterHand2.6M, but we re-train and test the model on Haggling. For the single hand approaches, we crop each individual hand from the dataset using the provided ground truth bounding box for re-training and testing.

	InterHand2.6M [18]		Haggling [12]
	MPJPE	shape-err	MPJPE
Zimmermann *et al.* [42]	36.364	-	22.735
Zhou *et al.* [41]	23.478	23.892	13.203
Boukhayma *et al.* [3]	16.925	17.984	42.924
Moon *et al.* [18]	16.888	-	22.735
Spurr *et al.* [31]	15.402	-	14.430
Ours	**13.071**	**10.398**	**11.419**

Table 1. Comparison of hand pose error (MPJPE) and shape error (shape-err) on InterHand2.6M and Haggling. Our method outperforms all the other methods. Since Haggling does not provide hand shape annotation, we do not compare the shape error on it.

Table 1 and Fig. 5 show the comparison on InterHand2.6M and Haggling. Our methods significantly outperform all the single hand approaches, presumably because they do not handle heavy hand occlusions. Compared to Moon *et al.* [18], our method also reduces hand MPJPE.

Secondly, we compare the hand shape performance to the state-of-the-art single hand shape reconstruction methods including Boukhayma *et al.* [3] and Zhou *et al.* [41]. Since the Haggling dataset does not contain the hand shape annotations, we only conduct the comparison of hand shape performance on the InterHand2.6M dataset. Our method also outperforms the compared single hand reconstruction methods by a huge margin on shape accuracy (See Table 1).

Fig. 6 further shows the qualitative comparison on InterHand2.6M [18]. Again, our method recovers significantly better hand pose and shape than the other methods. Fig. 7 shows qualitative results of interacting hand reconstruction with our network on InterHand2.6M and Haggling. More results can be found in the supplementary material.

4.3. Ablation Study

In order to investigate the contribution of the key components of our method, we conduct ablation study on Inter-

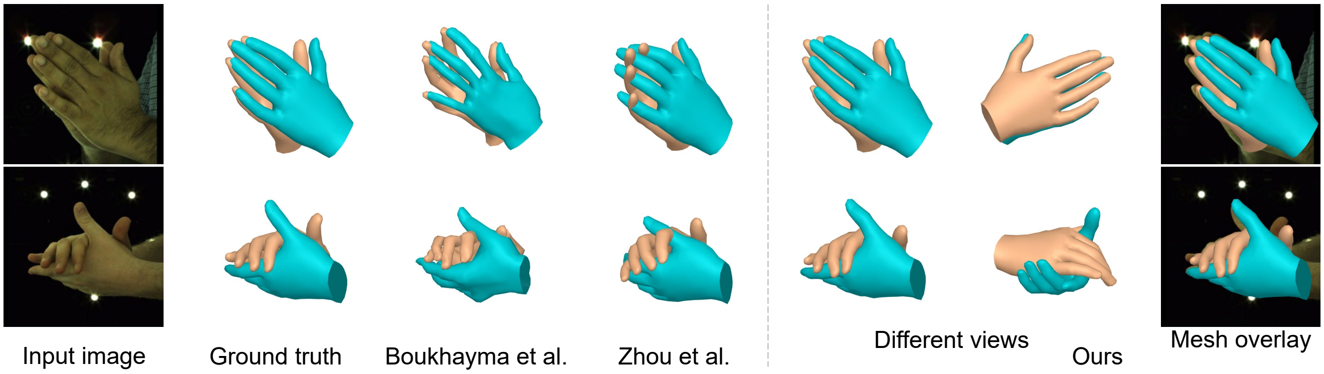


Figure 6. Qualitative comparison of the interacting hand reconstruction with our method and the state-of-the-art single hand reconstruction methods Boukhayma *et al*. [3] and Zhou *et al*. [41] on InterHand2.6M.

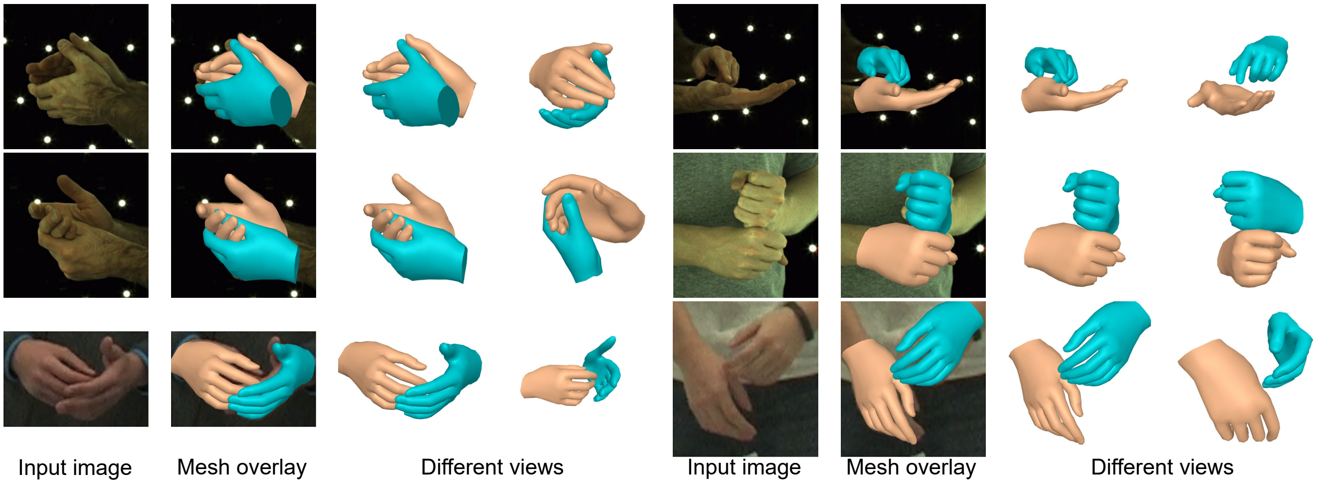


Figure 7. Qualitative results of interacting hand reconstruction with our network on InterHand2.6M (row1-row2) and Haggling (row3). Our method can achieve high-quality reconstruction performance under a variety of viewpoints and different levels of interhand occlusion.

Hand2.6M. By default, our full model means the full network in Fig. 2 with all loss functions and the complete network architecture. We compare with other methods to obtain the attention map, and investigate the effect of context-aware refinement and different inputs of the cascaded block.

Effect of Pose-aware Feature In our method, the predicted MANO parameters by the cascaded module is used to generate the attention map, and then extract pose-aware features. We compared different attention map generation approaches and show results in Fig. 8. 1) "Predicted 2.5D heatmap": Use predicted 2.5D heatmap [18] to generate attention map; 2) "Fit camera parameters": Use predicted the MANO parameters to generate 3D joints, render the joints to generate the heatmap, and generate the attention map with Eq. (3). 3) "No attention": Use the network architecture with cascaded blocks but without attention modules. 4) "Baseline": Use network architecture without cascaded blocks and attention modules.

We conduct comparison by removing the attention module in the cascaded refinement (Fig. 8, "no attention"). Experiments show that this method is inferior to the methods with attention using rendered heatmap or using predicted 2.5D attention map.

Experiments demonstrate that using attention map generated by rendering predicted joints is conducive to the improvement of accuracy, and the alignment method to calculate the camera parameters for rendering 3D joints (Fig. 8, "fit camera parameters") is more effective than the camera parameters predicted by a network (Fig. 8, "predict camera parameters"). However, the use of 2.5D heatmap to generate the attention map (Fig. 8, "predicted 2.5D heatmap") leads to a decrease in accuracy compared to our attention map. The main reason might be that the initial 2.5D heatmap prediction accuracy is not great, and the generated attention map cannot extract effective features for the cascaded blocks.

Effect of Context-aware Refinement To investigate the effectiveness of the context-aware refinement, we modify

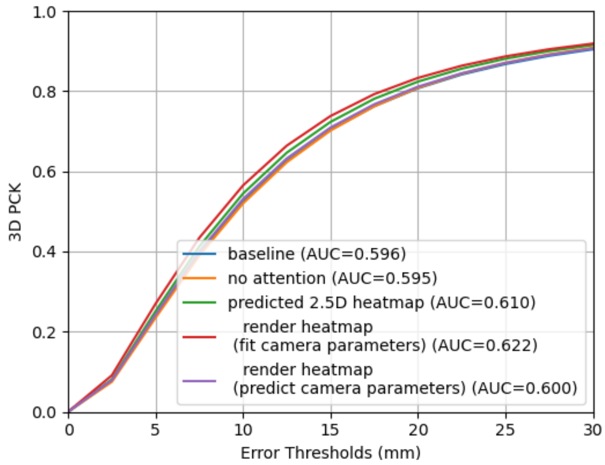


Figure 8. Comparison of different attention map generation methods. "Baseline": the network without cascaded block or attention module. "No attention": the network with cascaded block but no attention module. "Predicted 2.5D heatmap": the network that generates attention map from the predicted 2.5D heatmap. "Render heatmap": the network that renders the predicted 3D joints on the image to a heatmap and use it as an attention map. The weak perspective camera parameters to render 3D joints are obtained through alignment ("fit camera parameters") in our full model or network prediction ("predict camera parameters"), respectively.

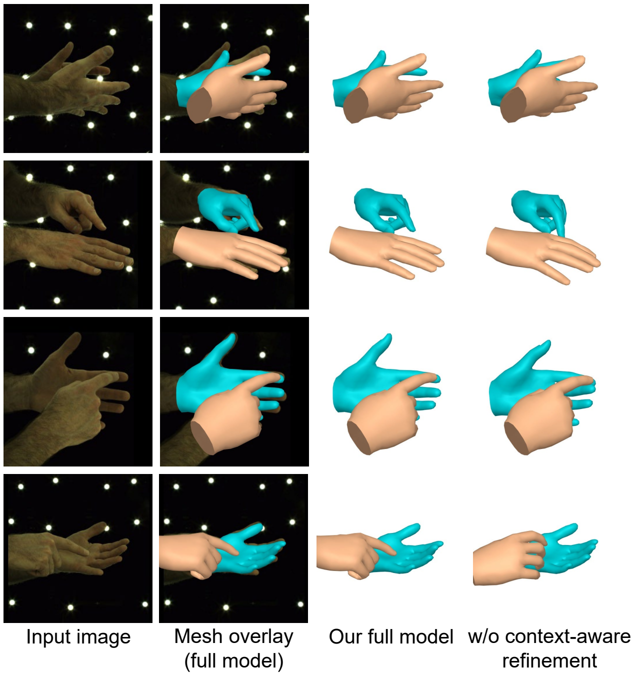


Figure 9. Qualitative study of context-aware refinement. We compare our full model and the model consisting cascaded blocks using MANO parameters of single hand as input.

the input of the cascaded blocks of our full model. Specifically, we use the predict MANO from the single hand that will be refined, instead of both hands, and the other inputs of the cascaded blocks are all remained. Table 2 shows that using the MANO parameters of two hands is better than just using one hand ("our full model" vs. "cascade single MANO parameters"). Although performance gain is relatively small, it is important for many applications. For example, tiny finger movement of interacting hands may lead to different interaction meanings, i.e. contact or separation of hands. Fig. 9 shows qualitative comparisons, and we find that our context-aware refinement can significantly improve the interacting two-hand reconstruction results.

	MPJPE	AUC(0-50mm)
baseline	14.218	0.734
no attention	14.095	0.735
predicted 2.5D heatmap	13.464	0.746
predict camera parameters	14.040	0.737
high level feature	13.986	0.737
cascaded single MANO parameters	13.170	0.752
our full model	**13.071**	**0.754**

Table 2. Ablation study of our network on InterHand2.6M. "High level feature" means that the attention map generated by the cascaded block always acts on the top feature of ResNet. "Cascaded single MANO parameters" means that the cascaded block for one hand only inputs the MANO parameters of this hand, which is predicted by the previous cascaded block, and the image feature of this hand. The other notations have the same meaning as Fig. 8.

Effect of Network Architecture To investigate the effect of different network architectures, we compare the performance of network architecture using the feature of the highest layer at the end of the encoder (Table 2, "high level feature") with our network using multi-scale features of lower layers of the encoder. We can observe that using multi-scale features is better than using high-level features ("our full model" vs. "high level feature").

5. Conclusion

In this work, we propose a novel solution to interacting hand pose and shape reconstruction. In order to address the key challenges of two-hand reconstruction, we propose a pose-aware attention module and context-aware cascaded refinement using two-hand correlation. The experiments demonstrate that our method can achieve the state-of-the-art interacting two-hand reconstruction performance on the main benchmark datasets. Our work can inspire related researches such as interacting hand reconstruction from video or depth, and whole-body reconstruction.

Acknowledgments This work was supported in part by National Natural Science Foundation of China (No. 62076061, No. 61473276, No. 61872346), Beijing Natural Science Foundation (4212029, L182052), and Newton Prize 2019 China Award (NP2PB/100047).

References

[1] 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*, 2020. 2

[2] Luca Ballan, Aparna Taneja, Juergen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Proceedings of the European Conference on Computer Vision*, 2012. 2

[3] Adnane Boukhayma, Rodrigo de Bem, and P. Torr. 3d hand shape and pose from images in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10835–10844, 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. *Proceedings of the European Conference on Computer Vision*, 2018. 2

[5] Xiaoming Deng, Yinda Zhang, Jian Shi, Yuying Zhu, Dachuan Cheng, Dexin Zuo, Zhaopeng Cui, Ping Tan, Liang Chang, and Hongan Wang. Hand pose understanding with large-scale photo-realistic rendering dataset. *IEEE Transactions on Image Processing*, 30:4275–4290, 2021. 2

[6] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2

[7] 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*, 2019. 2

[8] Shangchen Han, B. Liu, R. Cabezas, Christopher D. Twigg, P. Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Z. Wang, Asaf Nitzan, G. Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics*, 39:87, 2020. 1, 2

[9] 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*, 2020. 1, 2

[10] Yana Hasson, Gül 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*, 2019. 2

[11] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. *Proceedings of the European Conference on Computer Vision*, 2018. 2

[12] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 5, 6

[13] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8320–8329, 2018. 2

[14] Dominik Kulon, Riza Alp Güler, I. Kokkinos, M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4989–4999, 2020. 2

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

[16] Fanqing Lin, Connor Wilhelm, and Tony Martinez. Two-hand global 3d pose estimation using monocular rgb. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2373–2381, January 2021. 2, 5

[17] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *Proceedings of the European Conference on Computer Vision*, 2020. 2

[18] Gyeongsik Moon, Shoou-I Yu, H. Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. *Proceedings of the European Conference on Computer Vision*, 2020. 1, 2, 3, 4, 5, 6, 7

[19] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics*, 38(4), 2019. 1, 2

[20] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, pages 483–499, 2016. 3

[21] M. Oberweger, P. Wohlhart, and V. Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(8):1898–1912, 2020. 2

[22] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the articulated motion of two strongly interacting hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1862–1869, 2012. 2

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pages 8024–8035. 2019. 5

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas,

and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[25] Mark Richardson, Matt Durasoff, and Robert Wang. Decoding surface touch typing from hand-tracking. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 686–696, 2020. 1

[26] Grégory Rogez, Jonathan Rihan, Srikumar Ramalingam, Carlos Orrite, and Philip HS Torr. Randomized trees for human pose detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2008. 4

[27] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 2, 3, 5

[28] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 2

[29] 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*, 39(6):1–14, 2020. 2

[30] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *Proceedings of the European Conference on Computer Vision*, volume 12362, pages 211–228, 2020. 2

[31] A. Spurr, J. Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 2, 6

[32] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, Arran Topalian, Erroll Wood, Sameh Khamis, Pushmeet Kohli, Toby Sharp, Shahram Izadi, Richard Banks, Andrew Fitzgibbon, and Jamie Shotton. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics*, 35(4):1–12, 2016. 2

[33] Dimitrios Tzionas, L. Ballan, A. Srikantha, Pablo Aponte, M. Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118:172–193, 2016. 2

[34] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, pages 2579–2605, 2008. 4

[35] 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*, 39(6):1–16, 2020. 2, 5

[36] Robert Wang, Sylvain Paris, and Jovan Popović. 6d hands: markerless hand-tracking for computer aided design. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*, pages 549–558, 2011. 1

[37] Yangang Wang, Baowen Zhang, and Cong Peng. Srhandnet: Real-time 2d hand pose estimation with simultaneous region localization. *IEEE Transactions on Image Processing*, 29:2977–2986, 2019. 3

[38] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand:rgb-sequence-based 3d hand pose and shape estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2020. 2

[39] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020. 1, 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2354–2364, 2019. 2

[41] Yuxiao Zhou, Marc Habermann, Weipeng Xu, I. Habibie, Christian. Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5345–5354, 2020. 1, 2, 6, 7

[42] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2, 6

[43] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russel, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2