ACR: Attention Collaboration-based Regressor for Arbitrary Two-Hand Reconstruction

Zhengdi Yu[1,2], Shaoli Huang[1]*, Chen Fang[1], Toby P. Breckon [2], Jue Wang
[1]Tencent AI Lab [2]Durham University
{zhengdiyu,shaolihuang,fcfang}@tencent.com toby.breckon@durham.com arphid@gmail.com

Abstract

Reconstructing two hands from monocular RGB images is challenging due to frequent occlusion and mutual confusion. Existing methods mainly learn an entangled representation to encode two interacting hands, which are incredibly fragile to impaired interaction, such as truncated hands, separate hands, or external occlusion. This paper presents ACR (Attention Collaboration-based Regressor), which makes the first attempt to reconstruct hands in arbitrary scenarios. To achieve this, ACR explicitly mitigates interdependencies between hands and between parts by leveraging center and part-based attention for feature extraction. However, reducing interdependence helps release the input constraint while weakening the mutual reasoning about reconstructing the interacting hands. Thus, based on center attention, ACR also learns cross-hand prior that handle the interacting hands better. We evaluate our method on various types of hand reconstruction datasets. Our method significantly outperforms the best interacting-hand approaches on the InterHand2.6M dataset while yielding comparable performance with the state-of-the-art single-hand methods on the FreiHand dataset. More qualitative results on in-the-wild and hand-object interaction datasets and web images/videos further demonstrate the effectiveness of our approach for arbitrary hand reconstruction. Our code is available at this link [1].

1. Introduction

3D hand pose and shape reconstruction based on a single RGB camera plays an essential role in various emerging applications, such as augmented and virtual reality (AR/VR), human-computer interaction, 3D character animation for movies and games, etc. However, this task is highly challenging due to limited labeled data, occlusion, depth ambiguity, etc. Earlier attempts [1, 2, 36, 39] level down the problem difficulty and focus on single-hand reconstruction. These methods started from exploring weakly-supervised learning paradigms [2] to designing more advanced network models [31]. Although single-hand approaches can be extended to reconstruct two hands, they generally ignore the inter-occlusion and confusion issues, thus failing to handle two interacting hands.

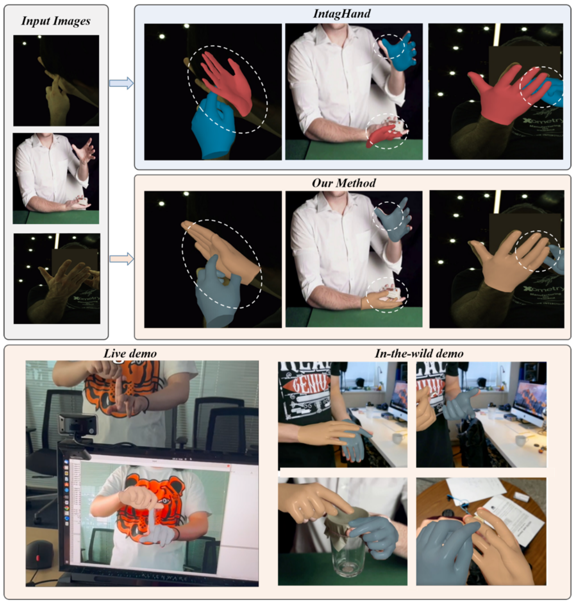

Figure 1. Given a monocular RGB image, our method makes the first attempt to reconstruct hands under arbitrary scenarios by representation disentanglement and interaction mutual reasoning while the previous state-of-the-art method IntagHand [15] failed.

To this end, recent research has shifted toward reconstructing two interacting hands. Wang et al. [33] extract multi-source complementary information to reconstruct two interacting hands simultaneously. Rong et al. [27] and Zhang et al. [35] first obtain initial prediction and stack intermediate results together to refine two-hand reconstruction. The latest work [15] gathers pyramid features and two-hand features as input for a GCN-based network that regresses two interacting hands unitedly. These methods share the same principle: treating two hands as an integral

*Corresponding author.

[1] https://github.com/ZhengdiYu/Arbitrary-Hands-3D-Reconstruction

and learning a unified feature to ultimately refine or regress the interacting-hand model. The strategy delivers the advantage of explicitly capturing the hands' correlation but inevitably introduces the input constraint of two hands. This limitation also makes the methods particularly vulnerable and easily fail to handle inputs containing imperfect hand interactions, including truncation or external occlusions.

This paper takes the first step toward reconstructing two hands in arbitrary scenarios. Our first key insight is leveraging center and part attention to mitigate interdependencies between hands and between parts to release the input constraint and eliminate the prediction sensitivity to a small occluded or truncated part. To this end, we propose Attention Collaboration-based Regressor (ACR). Specifically, it comprises two essential ingredients: Attention Encoder (AE) and Attention Collaboration-based Feature Aggregator (ACFA). The former learns hand-center and per-part attention maps with a cross-hand prior map, allowing the network to know the visibility of both hands and each part before the hand regression. The latter exploits the hand-center and per-part attention to extract global and local features as a collaborative representation for regressing each hand independently and subsequently enhance the interaction modeling by cross-hand prior reasoning with an interaction field. In contrast to the existing method, our method provides more advantages, such as hand detector free. Furthermore, experiments show that ACR achieves lower error on the InterHand2.6M dataset than the state-of-the-art interacting-hand methods, demonstrating its effectiveness in handling interaction challenges. Finally, results on in-the-wild images or video demos indicate that our approach is promising for real-world application with the powerful aggregated representation for arbitrary hands reconstruction.

Our key contributions are summarized as: ***(1)*** we take **the first step toward reconstructing two hands at arbitrary scenarios**. ***(2)*** We propose to **leverage both center and part based representation to mitigate interdependencies between hands and between parts** and release the input constraint. ***(3)*** In terms of modeling for interacting hands, we **propose a cross-hand prior reasoning module with an interaction field** to adjust the dependency strength. ***(4)*** Our method **outperforms existing state-of-the-art approaches significantly on the InterHand2.6M benchmark**. Furthermore, ACR is the most practical method for various in-the-wild application scenes among all the prior arts of hand reconstruction.

2. Related Work

Single-Hand Reconstruction: Hand pose and shape reconstruction from monocular images has rapidly progressed thanks to the development of the 3D hand parameterized model (e.g., MANO [26] and DeepHandMesh [22]). However, hand-mesh annotations are expensive and difficult to acquire, which constitutes the main obstacle for this task. Existing works [1, 2, 36, 39] tackled the issue mainly by exploiting weakly-supervised learning paradigms or synthesizing pseudo data. For example, Boukhayma et al. [2] utilized 2D/3D keypoints as weak supervision to guide MANO parameter regression. Zhang et al. [36] and Baek et al. [1] introduced segmentation masks as extra weak labels in training by employing a neural mesh renderer [12]. Rather than using standard 2D labels, Zhou et al. [39] leveraged motion capture data for weak supervision and proposed an inverse kinematics network to recover hand mesh from 3D keypoints. Generating pseudo data is another effective way to mitigate mesh-label scarcity. Kulon et al. [14] adopted a parametric model-fitting approach to generate pseudo mesh ground truth, enabling fully-supervised training for mesh reconstruction. Ge et al. [7] created a synthetic dataset by blending a rendered hand with a background image and further trained a Graph CNN-based method with full supervision. Recently, with the emergence of new hand pose and shape datasets (e.g., FreiHAND [42]), the latest work focused on developing more advanced network models or learning strategies to improve reconstruction accuracy. For example, Moon and Lee [21] proposed an image-to-lixel network that considers prediction uncertainty and maintains the spatial relationship. In addition, Tang et al. [31] proposed decoupling the hand-mesh reconstruction task into multiple stages to ensure finer reconstruction. Though these approaches have steadily improved hand reconstruction from monocular images, they are dedicated to the solo hand and usually fail to work well on two-hand cases. In contrast, our method explicitly addresses the challenge of inter-hand occlusion and confusion and, therefore, can deal with two interacting hands.

Two-Hand Reconstruction: A straightforward way to deal with two-hand reconstruction is to locate each hand separately and then transform the task into single-hand reconstruction. This strategy is commonly adopted in full-body reconstruction frameworks [4, 6, 11, 34, 37, 40]. However, independently reconstructing two hands remains a failure in addressing interacting cases, as the closer hands usually inter-occlude and easily confuse the model prediction. Earlier works successfully dealt with hand interaction mainly relied on model fitting and multi-view or depth camera setup. For instance, Taylor et al. [32] introduced a two-view RGBD capture system and presented an implicit model of hand geometry to facilitate model optimization. Mueller et al. [24] simplified the system by only using a single depth camera. They further proposed a regression network to predict segmentation masks and vertex-to-pixel correspondences for pose and shape fitting. Smith et al. [28] adopted a multi-view RGB camera system to compute keypoints and 3D scans for mesh fitting. To handle self-interaction and occlusions, they introduced a physicallybased deformable model that improved the robustness of vision-based tracking algorithms.

Recent interest has shifted to two-hand reconstruction based on a single RGB camera. Wang et al. [33] proposed a multi-task CNN that predicts multi-source complementary information from RGB images to reconstruct two interacting hands. Rong et al. [27] introduced a two-stage framework that first obtained initial prediction and then performed factorized refinement to prevent producing colliding hands. Similarly, Zhang et al. [35] predicted the initial pose and shape from deeper features and gradually refined the regression with lower-layer features. The latest work [15] introduced a GCN-based mesh regression network that leveraged pyramid features and learned implicit attention to address occlusion and interaction issues. However, these methods primarily treat two hands as an integral and implicitly learn an entangled representation to encode two-hand interaction. In contrast, our approach learns independent features for each hand and exploits attention-conditioned cross-hand prior with local and global cues to address interacting challenges collaboratively.

3. Methodology

Unlike existing works [2, 5, 15, 20, 41] that rely on an external detector to perform entangled bounding-box-level representation learning. Fig. 2 presents the overview of our method ACR. Given a single RGB image **I** as input, ACR outputs 4 maps, which are Cross-hand Prior map, Parameter map, Hand Center map, and Part Segmentation map. Based on Parameter map, which predicts weak-perspective camera parameters and MANO parameters for both left hand and right hand at each pixel, ACR then leverages three types of pixel-level representations for attention aggregation from Parameter map. First, ACR explicitly mitigates inter-dependencies between hands and between parts by leveraging center and part-based representation for feature extraction using part-based attention. Moreover, ACR also learns a cross-hand prior for handling the interacting hands better with our third Cross-hand Prior map. Finally, after aggregating the representations, we feed estimated parameters $\boldsymbol{F_{out}}$ to MANO [26] model to generate hand meshes.

3.1. Preliminaries: Hand Mesh Representation

We use a parametric model MANO [26] to represent hand, which contains a pose parameter $\theta \in \mathbb{R}^{16\times3}$ and a shape parameter $\beta \in \mathbb{R}^{10}$. We utilize 6D representations [38] to present our hand pose as $\theta \in \mathbb{R}^{16\times6}$. The final hand mesh M could be reconstructed via a differentiable MANO model: $M = W(\beta, \theta)$. Subsequently, 3D joints $J_{3D} \in \mathbb{R}^{21\times3}$ can be retrieved from the mesh: $\hat{J_{3D}} = RM$, where R is a pre-trained linear regressor and $M \in \mathbb{R}^{778\times3}$.

3.2. Representations of Attention Encoder

In this section, we will present the details of each output map or AE (Attention Encoder) module and their representations as shown in Fig. 2. Given a monocular RGB image, we first extract a dense feature map $F \in \mathbb{R}^{C\times H\times W}$ through our CNN backbone. ACR then leverages three types of pixel-level representations for robust arbitrary hand representations disentanglement and mutual reasoning under complex interaction scenarios. For clarity, we denote the handedness by $\boldsymbol{h} \in \{L, R\}$.

Parameter map: $M_p \in \mathbb{R}^{218\times H\times W}$ can be divided into two maps for left hand and right hand separately, where the first 109 dimensions are used for left-hand feature aggregation and the rest for the right hand. For each of the map $M_p^h \in \mathbb{R}^{109\times H\times W}$. The 109 dimensions consist of two parts, MANO parameter $\theta \in \mathbb{R}^{16\times6}$, $\beta \in \mathbb{R}^{10}$ and a set of weak-perspective camera parameters (s, t_x, t_y) that represents the scale and translation for the 2D projection of the individual hand on the image. This map serves as our **base** module for aggregated representation learning.

Hand Center map: $A_c \in \mathbb{R}^{2\times H\times W}$ consists of two parts for left hand and right hand, which can be represented as $A_c^h \in \mathbb{R}^{1\times H\times W}$. Each of the maps is rendered as a 2D Gaussian heatmap, where each pixel represents the probability of a hand center being located at this 2D position. The center is defined as the center of all the visible **MCP** joints, the joints that connect fingers with palm. For adaptive global representation learning, we generate heatmaps by adjusting the Gaussian kernel size K according to the bounding box size of the hand in data preparation for supervision (details in Supplementary Material). As the **first** representation of ACR, this map explicitly mitigates inter-dependencies between hands and serves as an attention mask for better global representation learning.

Part Segmentation map: $A_p \in \mathbb{R}^{33\times H\times W}$ is learnt as a probabilistic segmentation volume. Each pixel on the volume is a channel of probability logits over 33 classes which consists of 1 background and 16 hand part classes for each hand corresponding to the MANO model. Thus we have $A_p^h \in \mathbb{R}^{16\times H\times W}$. We obtain the part segmentation mask obtained by rendering the ground truth MANO hand mesh using a differentiable neural renderer [12]. As the **second** representation of ACR, this map serves as an attention mask for part representation learning.

Cross-hand Prior map: $M_c \in \mathbb{R}^{218\times H\times W}$ contains two maps, $M_c^h \in \mathbb{R}^{109\times H\times W}$. It is split into two sets of parameters which are MANO parameter $\theta \in \mathbb{R}^{16\times6}$, $\beta \in \mathbb{R}^{10}$ and 3 camera parameters for cross hand **inverse** feature query. Empirically, the two hands' pose will be highly correlated when they are closely interacting within the interaction field (**IF**), which is introduced in 3.4. As our **third** representation, aggregating this module into our robustly disentangled representations is providing us with the

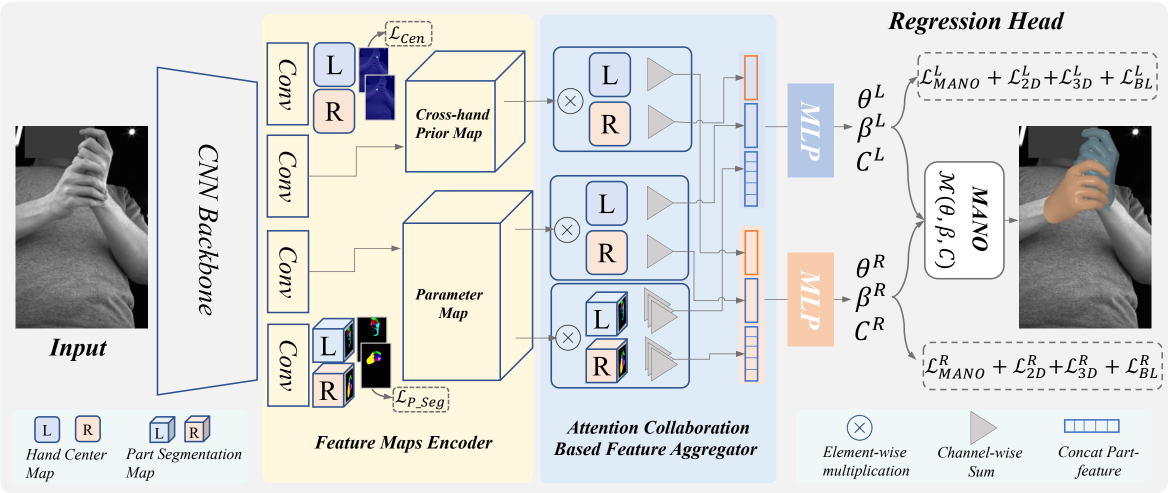
Input
CNN Backbone
Conv
Conv
Conv
Conv
L
R
$\mathcal{L}_{Cen}$
Cross-hand Prior Map
Parameter Map
$\mathcal{L}_{P\_Seg}$
Feature Maps Encoder
Attention Collaboration Based Feature Aggregator
MLP
MLP
Regression Head
θ^L
β^L
C^L
θ^R
β^R
C^R
$\mathcal{L}^L_{MANO} + \mathcal{L}^L_{2D} + \mathcal{L}^L_{3D} + \mathcal{L}^L_{BL}$
MANO $\mathcal{M}(\theta, \beta, C)$
$\mathcal{L}^R_{MANO} + \mathcal{L}^R_{2D} + \mathcal{L}^R_{3D} + \mathcal{L}^R_{BL}$
Hand Center Map
Part Segmentation Map
Element-wise multiplication
Channel-wise Sum
Concat Part-feature

Figure 2. **ACR network architecture:** ACR takes a full-person image and uses a feature map encoder to extract hand-center maps, part-segmentation maps, cross-hand prior maps, and parameter maps. Subsequently, the feature aggregator generates the final feature for the hand model regression based on these feature maps.

powerful mutual reasoning ability under severe interaction.

3.3. Robust Representation Disentanglement

Existing approaches for interacting hands reconstruction [15, 23, 35] typically require that the input image must be fixed to **two closely** interacting hands and occupy the most region of the image. This will cause ambiguity and unnecessary input constraints as shown in Fig. 1. In contrast, our first step towards building **arbitrary** hands representation is - **disentanglement** by decomposing the ambiguous hand representations. Thanks to the powerful pixel-wise representation of Hand Center map, we are able to disentangle **inter**-hand dependency and build an explicitly separate feature representation for the two hands. However, these feature representations could also be highly ambiguous when the two centres are getting closer. Subsequently, for better disentangled feature representation learning, inspired by [30], we adopt a collision-aware center-based representation to further split the features of two hands by applying Eq. 1. When the two hands are too close to each other with a Euclidean distance d smaller than $k_L + k_R + 1$. The new centers will be generated as follows:

$$\begin{aligned} \hat{\boldsymbol{C}_{\boldsymbol{L}}} &= C_L + \alpha R, \quad \hat{\boldsymbol{C}_{\boldsymbol{R}}} = C_R - \alpha R, \\ \boldsymbol{R} &= \frac{k_L + k_R + 1 - d}{d}(C_L - C_R) \end{aligned} \tag{1}$$

where C_L, k_L and C_R, k_R stand for two hand centers and their kernel size. R means the repulsion vector from C_L to C_R. In addition, α refers to an intensity coefficient to adjust the strength. Finally, the global representation $\boldsymbol{F^h_g} \in \mathbb{R}^{J*6+(10+3)}$, is extracted by combing Hand Center map A_c with parameter map M_p as:

$$F^h_g = f_g(\sigma(A^h_c) \otimes M^h_p) \tag{2}$$

where $\sigma, \odot$ and f_g are spatial softmax, pixel-wise multiply and a point-wise Multi-Layer Perceptron (MLP) layer separately, and $\boldsymbol{h} \in \{L, R\}$.

With such global feature representation F_g, we have disentangled inter-dependency. However, having only such global representation will lead to instability under occlusion and losing the ability to recover details, due to the unnecessary **inner** dependency of each hand part. Subsequently, we need to further disentangle our representation utilizing our Part Segmentation map A_p following [13]. For simplicity, we ignore the $\boldsymbol{h} \in \{L, R\}$ here, the two hands follow the same formulation as:

$$F^{(j,c)}_p = \sum_{h,w} \sigma(A^j_p) \odot M^c_p, \tag{3}$$

where $F_p \in \mathbb{R}^{J \times C}$ is final part representation and $F^{(j,c)}_p$ is its pixel at (j, c). $\odot$ is the Hadamard product. Thus, the part segmentation maps after spatial softmax normalization σ are used as soft attention masks to aggregate features in M^c_p. We follow prior arts to implement a dot product based method by reshaping the tensor at first: $F_p = \sigma(A^*_p)^T M^*_p$, where $M^*_p \in R^{HW \times C}$ and $A^*_p \in R^{HW \times J}$ are the parameter map M_p and reshaped part segmentation A_p without background mask. Finally, the global feature representation F_g and part representation and F_p are aggregated into our Robust Inter and Inner Disentangled Representation.

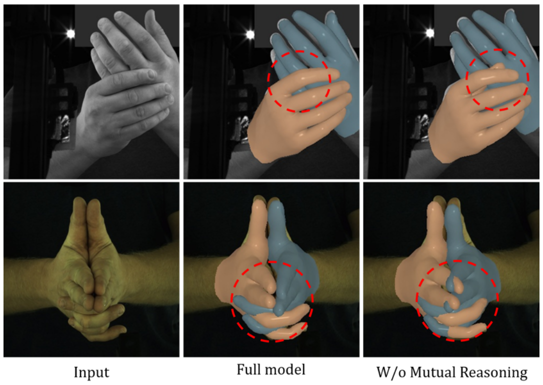

Figure 3. This compares our full model and a model without mutual reasoning, which explicitly helps deduce and recover the correlation between closely interacting hands with less confusion.

3.4. Mutual Reasoning of Interaction

Despite the powerful disentangled representations, it has been explored that the states of two interacting hands are highly correlated [15, 35] when they are interacting closely. Simply disentangling inter and inner dependencies as the final representation will weaken the mutual reasoning about reconstructing the interacting hands. Subsequently, we design a novel mutual reasoning strategy by reusing the center-based attention via a **inverse query**:

$$\begin{aligned} F_c^{R\rightarrow L} &= f_c(\sigma(A_c^R) \otimes M_c^L), \\ F_c^{L\rightarrow R} &= f_c(\sigma(A_c^L) \otimes M_c^R), \end{aligned} \tag{4}$$

where $F_c^{R\rightarrow L}$ is the left-hand prior representation that is deduced from right-hand attention and vice versa. M_c is the output dense feature map from cross-hand-prior attention blocks, A_c is our center based attention map, and L, R stand for left hand and right hand. σ, $\otimes$ and f_c are spatial softmax, pixel-wise multiply and a point-wise MLP layer.

However, for two more distant hands or a single hand, the correlation between them should be mitigated or eliminated. Subsequently, we also propose a new mechanism, interaction field (**IF**) to adjust the dependency strength. Specifically, by first computing the Euclidean distance $\boldsymbol{d}$ between the hands, when the two hands are too close to each other and entering the field of **IF**$= \gamma(k_L + k_R + 1)$, where γ is a field sensitivity scale, and the interaction intensity coefficient $\boldsymbol{\lambda}$ will be computed as:

$$\boldsymbol{\lambda}(\boldsymbol{C_L}, \boldsymbol{C_R}) = \begin{cases} 0, & d > IF \\ \frac{IF-d}{d}||C_L - C_R||_1, & d <= IF \end{cases}$$

The intensity coefficient $\boldsymbol{\lambda}$ helps our cross-hand prior representation to formulate an adaptive interaction field that can better model the correlations of two hands while keeping sensitive to close interaction and separation to avoid unnecessary feature entanglement. Finally, our final output self-adaptive robust representation could be represented as:

$$F_{out}^h = f_{out}(concat(F_g^h, F_p^{h*}, \lambda F_c^{\hat{h}\rightarrow h})) \tag{5}$$

where f_{out} is point-wise MLP layers for regressing the final representation $F_{out}^h \in \mathbb{R}^{109}$, and $F_c^{h*} \in \mathbb{R}^{J*C}$ is reshaped part disentangled representation. Finally, the results are fed into MANO model to regress the final hand mesh. For simplicity, we represent the opposite hand by $\hat{h}$ in Eq. 5.

3.5. Loss Functions

For training ACR with three types of powerful representation, our loss functions are divided into three groups, as demonstrated in Fig 2. Specifically, ACR is supervised by the weighted sum of all loss items for both left hand and right hand: mesh recovery loss, center-based attention loss, and part-based attention loss.

Center Attention Loss can be treated as a segmentation problem, however, the Gaussian distribution on the image is a relatively small area and there is an imbalance between the positive and negative samples. Subsequently, we utilize focal loss [18] to supervise our center map regressor as:

$$\boldsymbol{\mathcal{L}_c} = \sum_{h\in\{L,R\}} f(A_c^h, \hat{A_c^h}), \tag{6}$$

where f is focal loss [18], $h \in \{L, R\}$ means left hand and right hand, and $\hat{A_c^h}$ is the ground truth hand center map for hand type h. For simplicity, here we abbreviate the formulation of focal loss.

Part Attention Loss is used to supervise our Part-based Representation learning. We only supervise this loss with CrossEntropy loss in the first 2 epochs and continue to train with other losses until it converges.

$$\boldsymbol{\mathcal{L}_{seg}} = \frac{1}{HW}\sum_{h,w} CrossEntropy(\sigma(A_p^{hw}), A_p^{\hat{hw}}), \tag{7}$$

where $\hat{A_p}$ means GT part segmentation maps and $A_p^{\hat{hw}}$ is the ground truth class label at the location of (h,w). Different from our part soft attention mask, $A_p^{hw} \in \mathbb{R}^{33\times1\times1}$ here means the probabilistic segmentation volume at the pixel position of (h, w) and σ means softmax along channel dimension. We do not need to omit the background class here.

Mesh Recovery Loss is applied for each hand, thus we ignore the handedness $\boldsymbol{h} \in \{L, R\}$ here for simplicity. Finally, the loss for left hand and right hand will be summed into the total loss. Instead of relying on the ground truth vertex positions, which could cause degeneration in generalization ability, we decouple our mesh loss into 3 parts:

$$\boldsymbol{\mathcal{L}_{mesh}} = \boldsymbol{\mathcal{L}_{mano}} + \boldsymbol{\mathcal{L}_{joint}}, \tag{8}$$

where $\mathcal{L}_{mano}$ is the weighted sum of $L2$ loss of the MANO parameters θ and β, namely $w_\theta \mathcal{L}_\theta + w_\beta \mathcal{L}_\beta$:

$$\boldsymbol{\mathcal{L}_{\theta}} = w_\theta||\theta - \hat{\theta}||_2^2, \quad \boldsymbol{\mathcal{L}_{\beta}} = w_\beta||\beta - \hat{\beta}||_2^2, \tag{9}$$

where $\mathcal{L}_{joint}$ is the weighted sum of $\mathcal{L}_{3D}$, $\mathcal{L}_{2D}$ and a bone length loss $\mathcal{L}_{bone}$ to present better geometric constraint to the reconstructed mesh, which is the $L2$ distance between i^{th} ground truth bone length $\hat{b_i}$ and predicted length b_i:

$$\begin{aligned} &\boldsymbol{\mathcal{L}_{3D}} = w_{j3d}\mathcal{L}_{MPJPE} + w_{paj3d}\mathcal{L}_{PA-MPJPE}, \\ &\boldsymbol{\mathcal{L}_{PJ2D}} = w_{pj2d}||PJ_{2D} - \hat{J_{2D}}||_2^2, \\ &\boldsymbol{\mathcal{L}_{bone}} = \sum_i ||b_i - \hat{b_i}||_2^2, \end{aligned} \tag{10}$$

where $\mathcal{L}_{MPJPE}$ is the $L2$ loss between ground-truth 3D joints $\hat{J_{3D}}$ and predicted ones J_{3D} retrieved from predicted mesh. $\mathcal{L}_{PA-MPJPE}$ is computed as the Procrustes-aligned mean per joint position error (PA-MPJPE). We do not supervise camera parameters directly, instead, the network adjusts the camera parameters by computing the $L2$ loss between ground truth $\hat{J_{2D}}$ and the projected 2d joints PJ_{2D} retrieved by a weak-perspective camera: PJ_{2D} as $x_{pj2d} = sx_{3D} + t_x, y_{pj2d} = sy_{3d} + t_y$. Finally, to compute $\mathcal{L}_{mesh}$ as a weighted sum, we apply $w_{j3d} = 200$, $w_{paj3d} = 360$, $w_{pj2d} = 400$, $w_{bl} = 200$. For $\mathcal{L}_{mano}$, we use $w_{pose} = 80$, $w_{shape} = 10$ in our experiments.

Total Loss is the weighted sum of the described loss above and can be represented as:

$$\boldsymbol{\mathcal{L}_{total}} = \boldsymbol{\mathcal{L}_{mesh}} + w_c\boldsymbol{\mathcal{L}_{c}} + w_p\boldsymbol{\mathcal{L}_{seg}}, \tag{11}$$

where $w_c = 80$, $w_p = 160$ and $\mathcal{L}_{mesh}$ is already a weighted sum. Each part is activated only when the corresponding ground truth is available. Finally, all of these losses are trained simultaneously in an end-to-end manner.

4. Experiments

Implementation details: We implement our network based on PyTorch [25]. For the backbone network, we have trained with both ResNet-50 [10] and HRNet-W32 [3], for faster inference speed or better reconstruction results respectively. Unlike existing approaches that require a hand detector, our method can reconstruct arbitrary hands in an end-to-end manner. Furthermore, our method does not limit its input to two-hand. Given a monocular raw RGB image without cropping or detection, all the input raw images and segmentation maps are resized to 512×512 while keeping the same aspect ratio with zero padding, then we extract the feature maps $f \in R^{(C+2) \times H \times W}$ from the backbone network with CoordConv [19]. The feature maps are finally fed into four Conv blocks to produce the four maps.

Training: For comparison on InterHand2.6M dataset, we train our model using Adam optimizer with a learning rate 5e-5 for eight epochs. We do not supervise L_{seg} and L_{MANO} when there is no MANO label valid because our ground truth segmentation is obtained from rendering ground truth MANO hand mesh using a neural renderer [12]. For all of our experiments, we initialized our network using the pre-trained backbone of HRNet-32W from [5] to speed up the training process. We train our network using 2 V100 GPUs with $batchsize$ of 64. The size of our backbone feature is 128×128 and the size of our 4 pixel-aligned output maps is 64×64. We applied random scale, rotation, flip, and colour jitter augmentation during training.

Testing: For all the experiments, if not specified, the backbone is HRNet-32W. For comparison with state-of-the-art, we use the full official test set for evaluation. The confidence threshold is set to 0.25 with a max detection number of one left hand and one right hand, as we only have one left hand and one right hand in all the training and testing sets.

Evaluation Metrics: To evaluate the accuracy of the two-hand reconstruction, we first report the mean per joint position error (MPJPE) and the Procrustes-aligned mean per joint position error (PA-MPJPE) in millimetres. Both errors are computed after joint root alignment following prior arts. We also studied the reconstruction accuracy of handshape by mean per-vertex position error (MPVPE) and the Procrustes-aligned mean per-vertex position error (PA-MPVPE) on the FreiHand dataset. Please see details of the metrics in supplementary materials.

4.1. Datasets

InterHand2.6M [23] is the only publicly available dataset for two-hand interaction with accurate two-hand mesh annotations. This large-scale real-captured dataset, with both accurate human (H) and machine(M) 3D pose and mesh annotation, contains 1.36M frames for training and 850K frames for testing. These subsets are split into two parts: interacting hands (IH) and single hand (SH). We use the 5 FPS IH subset with H+M annotations.

FreiHand and **HO-3D** dataset. FreiHand [42] is a single hand 3D pose estimation dataset with MANO annotation for each frame. It has 4×32,560 frames for training and 3960 frames for evaluation and testing. HO-3D [8] is a hand-object interaction dataset that contains 66K training images and 11K test images from a total of 68 sequences.

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

Results on InterHand2.6M dataset: We first compare our method with single-hand and interacting-hand approaches on InterHand2.6M. We follow the official split to train our model, and we report results on the official test split of the InterHand2.6M dataset for a fair comparison. As the reported result of IntagHand and [35] are obtained from a fil-

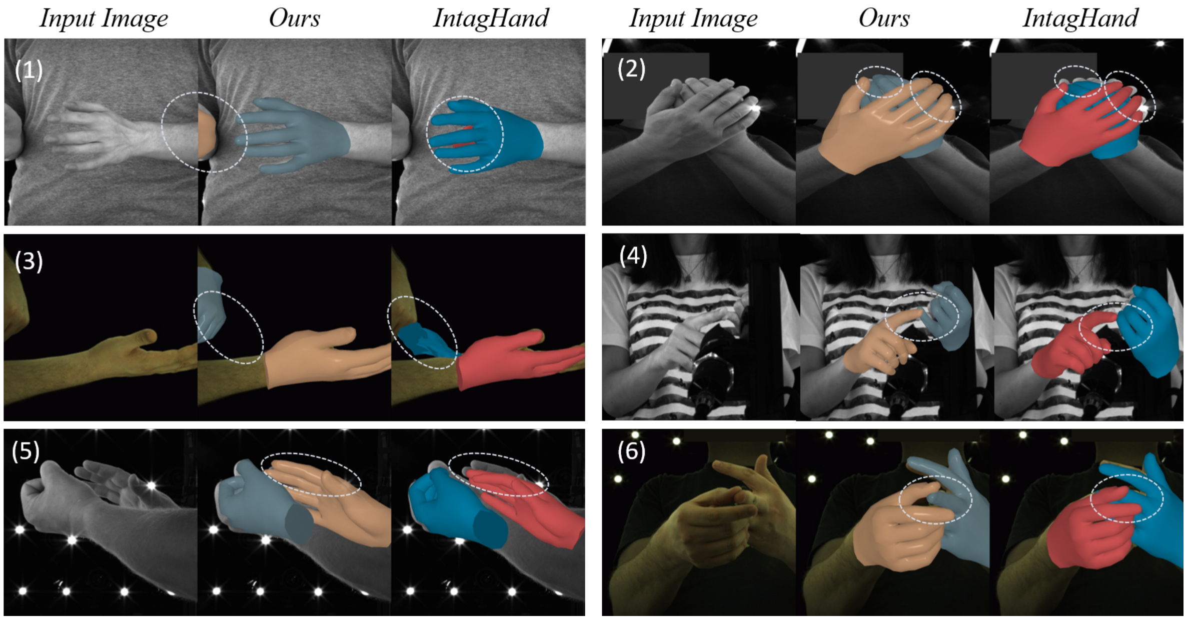

Figure 4. Qualitative comparison with on InterHand 2.6M test dataset. Our approach generates better results in two-hand reconstruction, particularly in challenging cases such as external occlusion (1), truncation (3-4), or bending one finger with another hand (6). More results can be found in the Supplementary Material.

	extra info.	MPJPE	MPVPE	IH MPJPE	IH MPVPE	SH MPJPE	SH MPVPE
(-) Zimmermann et al.[41]	Box	-	-	36.36	-	-	-
(-) Zhou et al.[39]	Box	-	-	23.48	23.89	-	-
(-) Boukhayma et al.[2]	Box	-	-	16.93	17.96	-	-
(-) Spurr et al. [29]	Box	-	-	15.40	-	-	-
Moon et al. [23]	Box	13.98	-	16.02	-	12.16	-
Fan et al. [5]	Box	-	-	14.27	-	11.32	-
IntagHand [15]	Box	9.95	10.29	10.27	10.53	9.67	9.91
Ours	**-**	**8.09**	**8.29**	**9.08**	**9.31**	**6.85**	**7.01**
Zhang et al. [35]*	Box+scale	11.58	12.04	11.28	12.01	11.73	12.06
IntagHand [15]*	Box+scale	9.18	9.42	9.40	9.68	9.0	9.18
Ours	scale	**7.41**	**7.63**	**8.41**	**8.53**	**6.09**	**6.21**

Table 1. Comparison with state-of-the-art on InterHand2.6M[23]. (-) means single hand reconstruction method. Except for our approach, all the others use ground-truth bounding boxes from the dataset. The single-hand results are taken from [35]. We report results on the official test split of the InterHand2.6M dataset for a fair comparison. * means the results are obtained by evaluating their released model on the official test split.

tered test set, we get the result on the standard test set by running their released code. Tab. 1 presents comparison results on the **I**nteracting hands (IH MPJPE), and **S**ingle hand (SH MPJPE) subset, and the full-set (MPJPE). Not surprisingly, we can observe that single-hand methods generally perform poorly on the IH subset, as their method designs dedicate to single-hand input. Next, we perform a comparison with two state-of-the-art interacting-hand approaches [35] and [15]. The first one adopted a refinement strategy that predicted the initial pose and shape from deeper features and gradually refined the regression with lower-layer features. The latter IntagHand incorporates pyramid features with GCN-based to learn implicit attention to address occlusion and interaction issues, while IntagHand is our concurrent work and outperforms [35]. However, our proposed method constantly surpasses IntagHand without extra information needed. Specifically, our method obtained the lowest MPJPE of 8.41 on the IH subset, demonstrating its effectiveness in handling interacting hands. It also achieves a 6.09 MPJPE on the SH dataset that outperforms IntagHand by a large margin, showing our method remains superior on single-hand reconstruction.

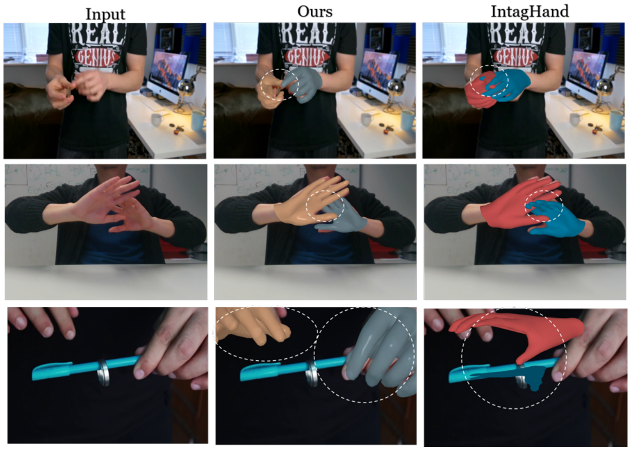

Figure 5. Qualitative comparison with IntagHand [15] on in-the-wild images.

Method	PA-MPJPE	PA-MPVPE
Mesh Graphormer[17]	6	5.9
METRO[16]	6.8	6.7
I2L-MeshNet[21]	7.4	7.6
HandTailor[20]	8.2	8.7
ours	6.9	7.0

Table 2. Comparison with state-of-the-art on FreiHand [42].

Method	Training Data	Crop	Joint Error ↓	AUC ↑
Keypoint Transformer [9]	HO3D	Yes	2.57	0.54
ACR(Ours)	InterHand2.6M+FreiHand	No	**2.14**	**0.61**

Table 3. Comparison with state-of-the-art on HO-3D [8].

Results on FreiHand and HO-3D dataset: We also compare our method with single-hand methods on the single-hand dataset FreiHand [42]. We follow the official split to train and test our model on this dataset. As in Tab. 2, the transformer-based method achieves the best result. Nevertheless, our method obtains comparable performance to the state-of-the-art single-hand approach, revealing its potential to improve single-hand reconstruction. Moreover, we test our model on HO-3D v2 dataset to demonstrate the ability to handle hand-object occlusion and truncation. In Tab. 3, we report the mean joint error after scale-translation alignment and the Area Under the Curve (AUC), where we can achieve 2.14 and 0.61 separately against [9] by only performing generalization.

	MPJPE	IH MPJPE	SH MPJPE	PAMPJPE
G(ResNet-50)	9.78	10.56	8.77	6.56
G(HRNet-32W)	9.56	10.35	8.65	6.41
P	8.70	9.76	7.26	5.59
G+C	9.1	9.88	8.11	6.08
G+P	8.52	9.69	6.87	5.49
G+C+P	**8.09**	**9.08**	**6.85**	**5.21**

Table 4. Ablation study on the part (P), global (G), and cross-hand (C) prior representation. We do not use any extra information such as bounding box and ground truth scale.

Qualitative Evaluation: We previously demonstrated our method significantly outperforms IntagHand in quantitative experiments. To gain insight into this result, we conduct a qualitative comparison between these two methods. Interestingly, our approach generally produces better reconstruction results over IntagHand in challenging cases such as external occlusion and truncated hands. Fig. 4 shows some examples of these cases. This result indicates that our method for two-hand reconstruction is less sensitive to some impaired observation. We also try our method to reconstruct in-the-wild images containing cases including single hand, ego-view, hand-object interaction and truncated hands. Fig. 5 presents some representative images where hands are accurately reconstructed, proving that our method has strong generality and is very promising for real-world applications.

4.3. Ablation study

As introduced in Sec. 3, our Attention Collaboration-based Feature Aggregator (ACFA) works mainly by collaborating three representations: **G**lobal representation (G, baseline), **P**art-based representation (P), and cross-hand-attention prior (C). Therefore, we investigate the effectiveness of each module. We treat the center-based representation as a baseline and gradually add another module to see their improvement. As shown in Tab. 4, we can clearly observe both part-based and cross-hand significantly improve the baseline. More interestingly, the improvement of adding C on the IH dataset is more significant than that on the SH dataset. This demonstrates cross-hand-attention prior facilitates addressing interacting hand challenges.

5. Conclusion and Future Work

Conclusion: We present a simple yet effective arbitrary hand reconstruction approach considering more challenges such as interacting hands, truncated hands, and external occlusion from monocular RGB image. To this end, we propose to leverage center and part attention to mitigate interdependencies between hands and between parts to release the input constraint and eliminate the prediction's sensitivity to a small occluded or truncated part. Experiments show that our method is a promising solution, which can serve as a baseline to inspire more research on arbitrary hand pose and shape reconstruction.

Limitation & Future Work: Our major limitation is the lack of explicit solution for mesh collision, resulting in occasional inter-penetration, which can be solved by leveraging relative information or perspective camera model for accurate depth reasoning and better simulation of translation.

References

[1] S. Baek, K. I. Kim, and T.-K. 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. 1, 2

[2] A. Boukhayma, R. d. Bem, and P. H. 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, 3, 7

[3] B. Cheng, B. Xiao, J. Wang, H. Shi, T. S. Huang, and L. Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020. 6

[4] V. Choutas, G. Pavlakos, T. Bolkart, D. Tzionas, and M. J. Black. Monocular expressive body regression through body-driven attention. In *European Conference on Computer Vision*, pages 20–40. Springer, 2020. 2

[5] Z. Fan, A. Spurr, M. Kocabas, S. Tang, M. Black, and O. Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *International Conference on 3D Vision (3DV)*, 2021. 3, 6, 7

[6] Y. Feng, V. Choutas, T. Bolkart, D. Tzionas, and M. J. Black. Collaborative regression of expressive bodies using moderation. In *2021 International Conference on 3D Vision (3DV)*, pages 792–804. IEEE, 2021. 2

[7] L. Ge, Z. Ren, Y. Li, Z. Xue, Y. Wang, J. Cai, and J. 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. 2

[8] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 6, 8

[9] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *CVPR*, 2022. 8

[10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[11] H. Joo, T. Simon, and Y. Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8320–8329, 2018. 2

[12] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 2, 3, 6

[13] M. Kocabas, C.-H. P. Huang, O. Hilliges, and M. J. Black. Pare: Part attention regressor for 3d human body estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11127–11137, 2021. 4

[14] D. Kulon, R. A. Guler, I. Kokkinos, M. M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4990–5000, 2020. 2

[15] M. Li, L. An, H. Zhang, L. Wu, F. Chen, T. Yu, and Y. Liu. Interacting attention graph for single image two-hand reconstruction. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 1, 3, 4, 5, 7, 8

[16] K. Lin, L. Wang, and Z. 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. 8

[17] K. Lin, L. Wang, and Z. Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. 8

[18] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[19] R. Liu, J. Lehman, P. Molino, F. Petroski Such, E. Frank, A. Sergeev, and J. Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. *Advances in neural information processing systems*, 31, 2018. 6

[20] J. Lv, W. Xu, L. Yang, S. Qian, C. Mao, and C. Lu. Handtailor: Towards high-precision monocular 3d hand recovery. *arXiv preprint arXiv:2102.09244*, 2021. 3, 8

[21] G. Moon and K. M. 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. 2, 8

[22] G. Moon, T. Shiratori, and K. M. Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*, pages 440–455. Springer, 2020. 2

[23] G. Moon, S.-I. Yu, H. Wen, T. Shiratori, and K. M. Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 4, 6, 7

[24] F. Mueller, M. Davis, F. Bernard, O. Sotnychenko, M. Verschoor, M. A. Otaduy, D. Casas, and C. Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Transactions on Graphics (TOG)*, 38(4):1–13, 2019. 2

[25] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[26] J. Romero, D. Tzionas, and M. J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6), 2017. 2, 3

[27] Y. Rong, J. Wang, Z. Liu, and C. C. Loy. Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In *2021 International Conference on 3D Vision (3DV)*, pages 432–441. IEEE, 2021. 1, 3

[28] B. Smith, C. Wu, H. Wen, P. Peluse, Y. Sheikh, J. K. Hodgins, and T. Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)*, 39(6):1–14, 2020. 2

[29] A. Spurr, J. Song, S. Park, and O. Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018. 7

[30] Y. Sun, Q. Bao, W. Liu, Y. Fu, B. Michael J., and T. Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 4

[31] X. Tang, T. Wang, and C.-W. Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11698–11707, 2021. 1, 2

[32] J. Taylor, V. Tankovich, D. Tang, C. Keskin, D. Kim, P. Davidson, A. Kowdle, and S. Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Transactions on Graphics (TOG)*, 36(6):1–12, 2017. 2

[33] J. Wang, F. Mueller, F. Bernard, S. Sorli, O. Sotnychenko, N. Qian, M. A. Otaduy, D. Casas, and C. Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 1, 3

[34] D. Xiang, H. Joo, and Y. Sheikh. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10965–10974, 2019. 2

[35] B. Zhang, Y. Wang, X. Deng, Y. Zhang, P. Tan, C. Ma, and H. 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. 1, 3, 4, 5, 6, 7

[36] X. Zhang, Q. Li, H. Mo, W. Zhang, and W. 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

[37] Y. Zhang, Z. Li, L. An, M. Li, T. Yu, and Y. Liu. Lightweight multi-person total motion capture using sparse multi-view cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5560–5569, 2021. 2

[38] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 3

[39] Y. Zhou, M. Habermann, W. Xu, I. Habibie, C. Theobalt, and F. 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, 7

[40] Y. Zhou, M. Habermann, I. Habibie, A. Tewari, C. Theobalt, and F. Xu. Monocular real-time full body capture with interpart correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4811–4822, 2021. 2

[41] C. Zimmermann and T. Brox. Learning to estimate 3d hand pose from single rgb images. Technical report, arXiv:1705.01389, 2017. URL `https://lmb.informatik.uni-freiburg.de/projects/hand3d/`. https://arxiv.org/abs/1705.01389. 3, 7

[42] C. Zimmermann, D. Ceylan, J. Yang, B. Russell, M. Argus, and T. 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*, pages 813–822, 2019. 2, 6, 8