Reconstructing Interacting Hands with Interaction Prior from Monocular Images

Binghui Zuo[1] Zimeng Zhao[1] Wenqian Sun[1] Wei Xie[1] Zhou Xue[2] Yangang Wang[1*]

[1]Southeast University, China [2]Pico IDL, ByteDance, Beijing

Abstract

Reconstructing interacting hands from monocular images is indispensable in AR/VR applications. Most existing solutions rely on the accurate localization of each skeleton joint. However, these methods tend to be unreliable due to the severe occlusion and confusing similarity among adjacent hand parts. This also defies human perception because humans can quickly imitate an interaction pattern without localizing all joints. Our key idea is to first construct a two-hand interaction prior and recast the interaction reconstruction task as the conditional sampling from the prior. To expand more interaction states, a large-scale multimodal dataset with physical plausibility is proposed. Then a VAE is trained to further condense these interaction patterns as latent codes in a prior distribution. When looking for image cues that contribute to interaction prior sampling, we propose the interaction adjacency heatmap (IAH). Compared with a joint-wise heatmap for localization, IAH assigns denser visible features to those invisible joints. Compared with an all-in-one visible heatmap, it provides more fine-grained local interaction information in each interaction region. Finally, the correlations between the extracted features and corresponding interaction codes are linked by the ViT module. Comprehensive evaluations on benchmark datasets have verified the effectiveness of this framework. The code and dataset are publicly available at https://github.com/binghui-z/InterPrior_pytorch.

1. Introduction

Reconstruction of interacting hands is significant for enhancing the behavioral realism of digital avatars in communication, thinking and working. With the advent of the RGB dataset [43] recording two-hand interactions, numerous attempts have been implemented to reconstruct interacting hands from monocular RGB images. Inspired by the existing single-hand frameworks [23, 69, 7], pioneer works [64, 12] localize and identify all two-hand joints as the interacting clues. Unfortunately, this process can be seriously misguided by the regional occlusion and local similarity between hands. Subsequent improvements include optimizing re-projection errors [51], localizing mesh vertices from coarse to fine [34], and querying all-in-one visible heatmap [27, 16]. Nevertheless, they still rely on more accurate joint 2D estimators, more diverse marker-less training data, or more computational complexity.

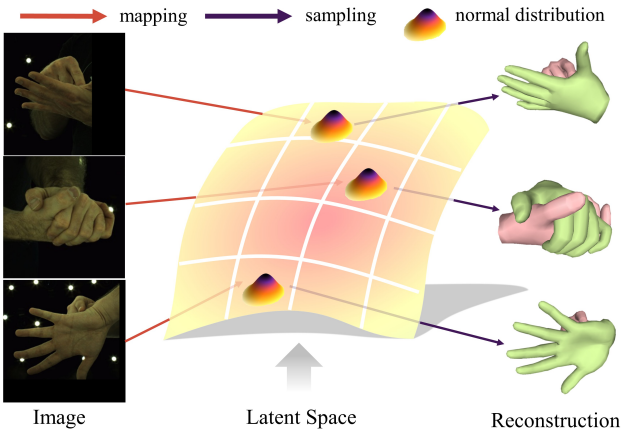


Figure 1. **Illustration of reconstructing interacting hands from monocular images by our framework**. *Left:* Using a ViT-based fusion network, we map the extracted features from inputs to the learned latent space. *Right:* We sample reasonable reconstructions from the pre-built interaction prior.

To overcome this hurdle, our key idea is to **first construct a comprehensive interaction prior with multimodal datasets and then sample this pre-built prior according to the interaction cues extracted from a monocular image.** It is noted that existing frameworks are al-

*Corresponding author. E-mail: yangangwang@seu.edu.cn. All the authors from Southeast University are affiliated with the Key Laboratory of Measurement and Control of Complex Systems of Engineering, Ministry of Education, Nanjing, China. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), in part by the Natural Science Foundation of Jiangsu Province (No. BK20220127).

ways trained with paired data of calibrated images and mesh annotations. This may lead to difficulty in generalizing since the well-known benchmark [43] contains simple backgrounds and only around 8.5K interaction patterns. We break this images-paired manner and construct an interaction prior with multimodal datasets, including marker-based data, marker-less data and hands-object data. To do this, a dataset with 500K two-hand patterns is proposed, which contains physically plausible 3D hand joints and MANO parameters. This dataset is used for the unsupervised training of a prior container, which can be formulated by a VAE [29]. As a result, each two-hand interaction pattern is mapped to an interaction code in the prior space. Since the correlation between the two hands is considered, this representation is more compact than doubling the hand joint/vertex positions or MANO parameters [49].

We argue that accurate joint localization is challenging for the monocular reconstruction of interacting hands. As an alternative, we sample the above pre-built interaction prior according to the *interaction adjacency heatmap* (IAH). This heatmap is defined as the mixture coordinate distribution of this joint and other two-hand joints within its coordinate neighborhood. Compared with the 2.5D joint heatmap [23, 64], our IAH abandons the pseudo depth and concatenates more on spatial correlations of the target joint. This heatmap formulation is easier to regress because even for an invisible joint, humans can determine its identity and location according to its spatial neighborhood. Considering that the Gaussian distribution has a more ambiguous boundary, the Laplacian distribution [2] is selected as the kernel function of each joint. This effectively reduces the aliasing of interacting adjacency information. These IAHs are further converted to be the corresponding interaction codes through the ViT [10] module and then are regarded as conditions to sample reasonable interaction from the latent space.

In summary, our main contributions are:

• A powerful interaction reconstruction framework that compactly represents two-hand patterns as latent codes, which are learned from multimodal datasets in an unsupervised manner.

• An effective feature extraction strategy that utilizes interaction adjacency as clues to identify each joint, which is inspired by human perception and is more friendly for network learning.

• A large-scale multimodal dataset that records 500K patterns of closely interacting hands, which is more conducive to the construction of our latent prior space.

2. Related Work

Hand monocular reconstruction has received a breakthrough after more than a decade of development. Previous studies have always focused on 3D joint estimation [6, 23, 44, 52, 71, 70, 11]. [70] introduced the first baseline to predict 3D joint position from a single image. The proposal of MANO [49] brings a new research direction to this field. With [5] presenting the first end-to-end solution for learning 3D hand shape and pose from a monocular RGB image, estimating model parameters [8, 69, 5, 67, 7, 66] from inputs has become a major trend. Some researches [13, 55, 31, 59] even directly regressed 3D vertices from the inputs. However, most of them are only applicable to a single-pose representation. In this work, we construct an interaction prior which is used to effectively estimate plausible hand poses. Our proposed unified framework can be applied to 3D joints, vertices and MANO parameters.

Interacting hand reconstruction is critical to promote the development of human-computer interaction(HCI). Due to the severe self-occlusion and the similar appearance of the entangled two hands, previous works heavily relied on depth cameras [45, 56, 57, 46] or multi-view cameras [4, 17]. Benefiting from the promotion of deep learning and the proposal of interacting hand dataset [43], previous works [34, 64, 51, 12, 16, 27, 39] tried to estimate interacting hand pose from monocular color images. Most of them attempted to extract distinguishable features of each hand [27, 12, 43, 60] or decouple the interaction [39, 50, 35]. Unfortunately, the traditional feature extraction schemes are unsuitable for extracting effective interaction details, and it is unreliable to decouple hands relying on features. Recently, an attention mechanism has been widely adopted [64, 34, 16] to yield more interacting attention features. Among them, Zhang *et al*. [64] utilized pose-aware attention and context-aware refinement module to improve the pose accuracy. Hampali *et al*. [16] employed a transformer architecture to model interaction, but the joint angle representation did not perform perfectly. Li *et al*. [34] fed the bundled features to the attention module, progressively regressing the two-hand vertices. Besides, [37, 36, 9] further demonstrated the power of the transformer. Inspired by the above studies, we propose a novel feature extraction module that gives more attention to the interaction region. Meanwhile, a ViT-based model is used to fuse them.

Learning-based prior has sparked more attention in different domains. Both VAE [29] and GAN [15] are the mainstream models for constructing diverse priors. Most related researchers constructed a prior to avoid implausible situations. To model human motion, [33, 48, 38, 26] used VAE to build motion prior. Other pioneers [18, 63, 22, 21, 47, 41] introduced pose or shape prior to refine geometric details. Similarly, [25, 30] designed adversarial prior to learn plausible reconstruction. Most similar to us are [53, 58, 61, 62], who applied the built prior to hand pose estimation. Wan *et al*. [58] combined GANs and VAEs to build two separate latent spaces. To estimate hand pose from depth maps, they used a mapping function to connect these two spaces. Spurr *et al*. [53] proposed a cross-modal framework where both

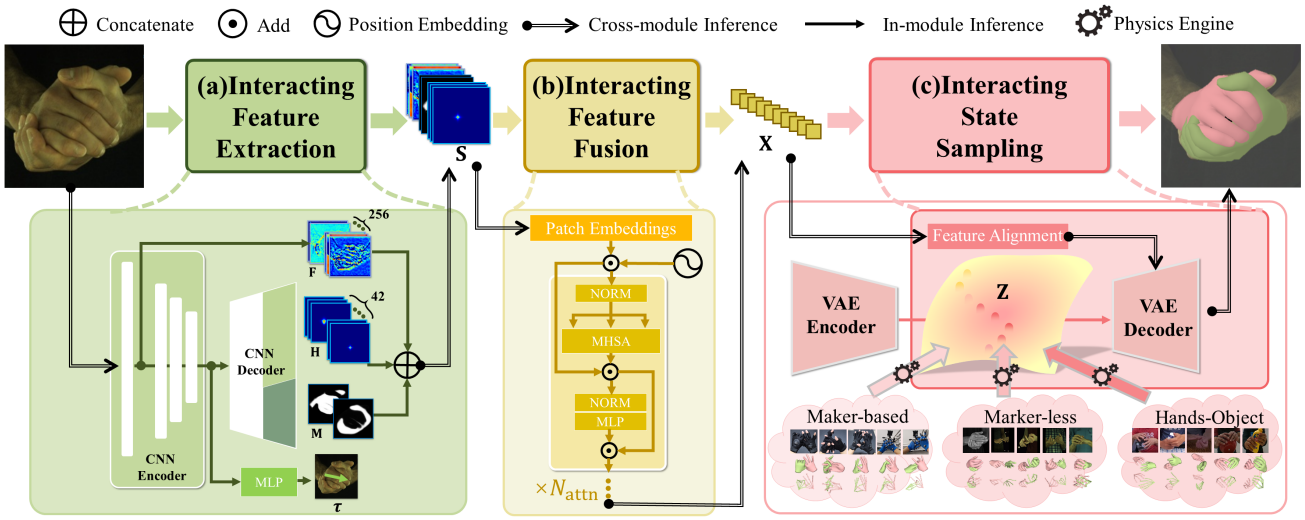


Figure 2. **Overview of our architecture**. The total pipeline consists of three stages. (a) We first design an expressive feature extraction module to extract global and local context information. Besides that, the proposed IAH adaptively maps adjacent joints with Laplacian distribution and provides more visual attention. (b) A ViT-based fusion module is designed to fuse the extracted features, which are regarded as sampling conditions to reconstruct the final result. (c) We build a powerful interaction prior with multimodal datasets and use the physics engine to make the ground truth more physically plausible.

RGB images and depth maps can be used to build the prior. In addition, [61, 62] committed to aligning the learned latent space with different modalities jointly. To this end, we extend the VAE framework to build a novel interaction prior. Compared with competitors, the biggest challenge of two-hand reconstruction is the lack of interaction status. Existing two-hand reconstruction methods are trained on [43] with the paired images. The smaller number of interaction states (8.5K) limits their generalization performance. Fortunately, our core improvement is to construct prior without paired images, which means multimodal datasets can be applied. To compensate for the lack of interaction between two hands, we provide a larger two-hand dataset *Two-hand 500K* , which has more interaction states than [43].

3. Method

The pipeline of the proposed method is shown in Fig. 2, which contains three stages: interacting feature extraction (Sec. 3.1), interacting feature fusion (Sec. 3.2) and interacting state sampling (Sec. 3.3). To facilitate the formulation, the hat superscripts represent the predicted result from the network, the star superscripts represent the ground truth values.

3.1. Interacting Feature Extraction

Feature selection. Given a monocular $\mathbf{I} \in \mathbb{R}^{(3,H,W)}$, we first extract the following 2D features related to the corresponding hand interaction: (i) Interaction adjacency heatmap (IAH) $\{\mathbf{H}_\mathrm{j}\}_{\mathrm{j}=1}^{42}$. (ii) Instance saliency maps $\mathbf{M} \triangleq \mathbf{M}_l \oplus \mathbf{M}_r$. (iii) Left-to-right relative translation $\boldsymbol{\tau} \in \mathbb{R}^3$.

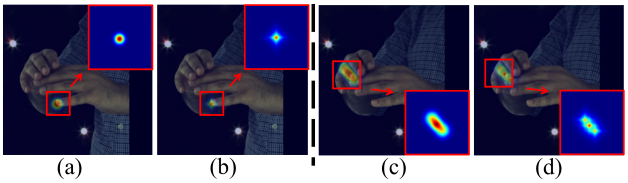


Figure 3. **IAHs of different channels**. (a) refers to $\mathbf{H}_{42}$ with Gaussian distribution, only the identity joint $\boldsymbol{\jmath}_{42}$ is mapped to the heatmap; (b) $\mathbf{H}_{42}$ with Laplacian distribution; (c) refers to $\mathbf{H}_{16}$, both $\boldsymbol{\jmath}_{16}$ and adjacent joints are mapped; (d) $\mathbf{H}_{16}$ with Laplacian distribution.

Because the joints of interacting hands are overcrowded and similar in $\mathbf{I}$, we argue that the strategy of simultaneous identification and localization[23, 64] is too tough for our task. As an alternative, our IAH emphasizes joint-wise identification more than localization. For a joint located at $\boldsymbol{\jmath}_\mathrm{j}(u_\mathrm{j}, v_\mathrm{j})$ and a $d \times d$ image adjacency $\mathcal{A}_\mathrm{j}(d)$, its IAH ground-truth is defined as a 2D Laplacian mixture distribution:

$$\mathbf{H}_\mathrm{j}^*(u,v) = \frac{1}{2\sigma_\mathrm{j}} \exp\left(-\frac{|u-u_\mathrm{j}| + |v-v_\mathrm{j}|}{\sigma_\mathrm{j}}\right) + \sum_{\boldsymbol{\jmath}_\mathrm{k} \in \mathcal{A}_\mathrm{j}(d)} \frac{1}{2(\alpha\sigma_\mathrm{j})} \exp\left(-\frac{|u-u_\mathrm{k}| + |v-v_\mathrm{k}|}{\alpha\sigma_\mathrm{j}}\right) \tag{1}$$

where the first term is the Laplacian kernel for the identity $\boldsymbol{\jmath}_\mathrm{j}$ with the variance σ_j. The other terms are the Laplacian kernels for adjacent joints $\boldsymbol{\jmath}_\mathrm{k} \in \mathcal{A}_\mathrm{j}(d)$ with the variance $\alpha\sigma_\mathrm{j}$. $\alpha > 1$ is a zoom factor. As shown in Fig. 3(c)&(d), we select Laplacian instead of Gaussian as the kernel function

mainly due to its clearer distribution boundary. $\mathbf{M}$ contains the visible parts of the left and right hand. Because $\boldsymbol{\tau}$ is often regarded as a global feature [64], it is regressed by an extra MLP. The overall loss term of this part is:

$$L_{2D} = \sum_{\mathrm{j}=1}^{42} \|\hat{\mathbf{H}}_{\mathrm{j}} - \mathbf{H}_{\mathrm{j}}^{\star}\|_2^2 + \lambda_1 \|\hat{\mathbf{M}} - \mathbf{M}^{\star}\|_2^2 + \lambda_2 \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^{*}\|_2^2 \tag{2}$$

In practice, ResNet50 [19] (with 4 cascading residual blocks) is selected as the feature extraction backbone, and the decoder for 2D local features is designed as a symmetrical structure with 4 blocks. For MLP used for regressing $\boldsymbol{\tau}$, after passing through an adaptive pooling layer behind the high-level feature map, we connect two fully connected layers to obtain $\boldsymbol{\tau}$.

Latent utilization. Besides the above explicitly supervised features, we further utilize the low-level feature maps $\mathbf{F}$ from the first block of our extractor as additional visual guidance. As shown in Fig. 2 (a), it contains more dense responses and the same map size as $\mathbf{H}$ and $\mathbf{M}$.

Implementation details. In our experiment, the variance σ_{j} of identity $\boldsymbol{\jmath}_{\mathrm{j}}$ is set to 2.0, the zoom factor α is set to 2.0 and adjacent region size d is set to 2.5. To balance each loss term, we set λ_1=1 and λ_2=2000. To normalize the translation, we fix the right translation to 0 and predict $\boldsymbol{\tau}$ between the left to right hand. We use Pytorch [3] to implement the feature extraction network and train it on a single NVIDIA GeForce RTX 3090. To update the network parameters, we use Adam [28] optimizer with a fixed learning rate 1e-4. We set the batch size to 64 and total training iterations to 500K. Before training, we crop the interacting hand regions with the annotated hand 2D vertices coordinates and resize it to 256×256. To improve the generalization, we perform data augmentation, including random rotation, random flip and color blur.

3.2. Interacting State Sampling

Prior construction. Building the prior allows us to sample the reasonable interaction from the extracted features even if one of the hands is completely occluded. Therefore, the expressiveness and accuracy of the constructed prior directly affect the final reconstruction. Similar to [53, 58, 61, 62], we deploy the VAE framework to build the interaction prior, which consists of an encoder and a decoder. The encoder implicitly maps the input $\boldsymbol{x}$ to $p(\boldsymbol{z})$ that conforms to the normal distribution, where $p(\boldsymbol{z})$ is the prior on the latent space. The decoder reconstructs $\hat{\boldsymbol{x}}$ that is close to $\boldsymbol{x}$. We represent the encoder as the conditional probability distribution $q(\boldsymbol{z} \mid \boldsymbol{x})$ and the decoder as $p(\hat{\boldsymbol{x}} \mid \boldsymbol{z})$. The building process is shown in Eqn. 3.

$$\boldsymbol{x} \overset{q(\boldsymbol{z}|\boldsymbol{x})}{\longrightarrow} p(\boldsymbol{z}) \overset{p(\hat{\boldsymbol{x}}|\boldsymbol{z})}{\longrightarrow} \hat{\boldsymbol{x}} \tag{3}$$

Training procedure. Both the encoder and decoder are composed of fully connected layers that follow ReLU activations. The encoder is a four-layers feed-forward network that models the input $\boldsymbol{x}$ to the latent space with dimension d_z. We force the output dimension of the encoder to $2d_z$, where the first d_z is used for the regression of mean μ and the second for variance σ. We strive to shape the distribution with μ and σ into a standard normal distribution and encourage it using Kullback-Leibler divergence loss. Afterward, the latent variable $\boldsymbol{z}$ sampled by the reparameterization trick is passed to the decoder to reconstruct the expected result $\hat{\boldsymbol{x}}$, where the decoder consists of six linear layers. We use MSE as the reconstruction loss to supervise it. The total loss for interaction prior is defined as:

$$L_{prior} = L_{KL}(\boldsymbol{z}) + \lambda_3 \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \tag{4}$$

In the inference phase, we discard the encoder and only use the decoder with the frozen parameters to obtain the reconstruction.

Hand representation. Benefiting from the embedding ability of VAE, we conduct experiments on three different hand pose representations, including 3D joints coordinates, 3D vertices coordinates and MANO parameters [49]. For compatibility with our framework, only pose and shape parameters are considered when embedding MANO parameters, as the relative translation has been estimated in the feature extraction module.

Implementation details. To balance quality and generalization, we set loss weight λ_3 to 100 to ensure losses are within one order of magnitude [38]. During training, we flatten inputs as a vector and map them to latent space with d_z=128. We use Adam [28] optimizer with a base learning rate of 1e-5 and a batch size of 64.

3.3. Interacting Feature Fusion

Feature fusion. To effectively leverage the features extracted in Sec. 3.1, a ViT-based fusion [10] module with powerful attention is used to fuse them. Specifically, the global image feature $\mathbf{F}$, interaction adjacency heatmap $\mathbf{H}$ and instance saliency maps $\mathbf{M}$ are concatenated to form fusion features $\mathbf{S} \in \mathbb{R}^{(C,H,W)}$. All of them have the same resolution H=W=64, and the channel C of $\mathbf{S}$ is 300. Before ViT, we reshape $\mathbf{S}$ into n visual patches, n=(HW/P^2), P denotes the size of each patch. We also use a single linear layer to map the patches into patch embeddings $\mathbf{S}'$ with dimension D, which is equal to the hidden size of all the transformer layers [10]. Extra position embeddings for preserving spatial information are added to the patch embeddings $\mathbf{S}'$. The fusion network follows a standard Transformer structure, which consists of multi-head self-attention layers (ATTE), feed-forward blocks (FF) and normalization layers (LN). We feed the embeddings into N_{attn} trans-

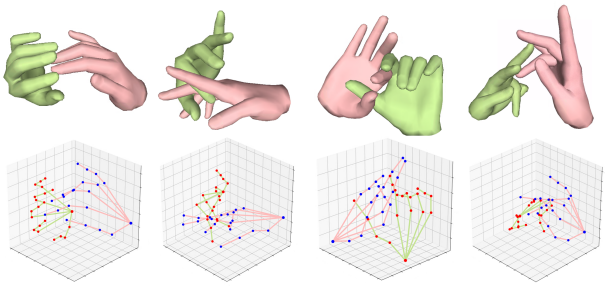


Figure 4. **The proposed *Two-hand 500K*** . Both 3D joint positions and corresponding MANO parameters are provided.

former attention blocks according to the following equation.

$$\begin{aligned} \tilde{\mathbf{S}}'_{n+1} &= \mathbf{S}'_n + \mathrm{ATTE}\left(\mathrm{LN}\left(\mathbf{S}'_n\right)\right), \\ \mathbf{S}'_{n+1} &= \tilde{\mathbf{S}}'_{n+1} + \mathrm{FF}\left(\mathrm{LN}\left(\tilde{\mathbf{S}}'_{n+1}\right)\right) \end{aligned} \tag{5}$$

where n denotes different transformer blocks and the tilde superscripts represent the intermediate output of the transformer.

Feature alignment. To ensure the consistency between the dimension of ViT output and the pre-built interaction prior, we add a feature alignment block consisting of a linear layer after the final transformer block. We treat the output of the feature alignment block as a condition and sample the expected reconstruction from the pre-built interaction prior. Only the VAE decoder with frozen parameters is employed for this process.

Training procedure. We train the ViT-based fusion and simultaneously fine-tune the feature extraction module in an end-to-end manner. In addition to the feature extraction loss defined in Eqn. 2, we also apply other special losses to supervise the reconstruction of different hand representations. For the representation of 3D joints and 3D vertices, we use the L_1 distance as a loss function to ensure consistency between the prediction and ground truth. And for the representation of MANO parameters, we also adopt additional loss terms to make the hand surface smoother and physically plausible, including normal loss [34, 42] and penetration loss [24, 51].

Implementation details. In our framework, we set the size of each patch P to 8, patch embeddings size D to 1024 and total use N_{attn}=6 transformer blocks. Different from feature extraction and interaction prior module, both the Adam [28] optimizer and the learning rate scheduler are used. The training process costs 150 epochs with a batch size of 64 on a single NVIDIA GeForce RTX 3090. We fix the learning rate at 1e-4 in the first 100 epochs and then adaptively reduce it with the scheduler. Besides the above, as the parameters of the feature extraction module are also updated, we use the same data augmentation as the training of the feature extraction network.

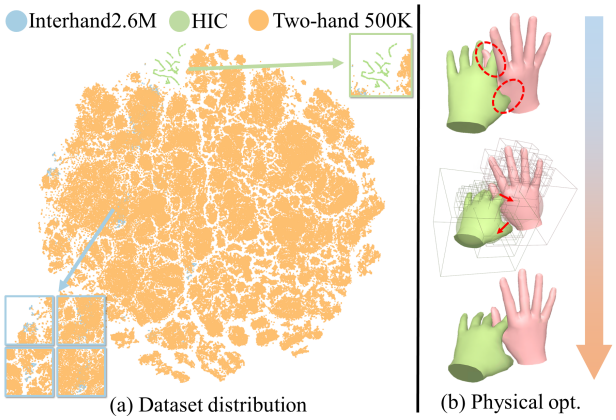


Figure 5. **Dataset distribution and physical optimization**. (a) t-SNE distribution of [43, 57] and our *Two-hand 500K* ; (b) physical optimization for the plausible interaction. From top to bottom: before optimization, during optimization, after optimization.

3.4. Interacting Modality Expansion

Motivation. To eliminate the dependence on image datasets when building the interaction prior, we propose the following measures to increase the diversity of hand interaction: (i) Capturing skeleton data according to the marker-based system simultaneously; (ii) Randomly combining the poses sampled from single-hand datasets. Fig. 4 shows our *Two-hand 500K* .

More diversity. To obtain more diverse interaction states, we construct *Two-hand 500K* in a multimodal manner. With the above data generation measures, more than 500K interaction states are captured. Besides the data captured by the marker-based MoCap system, we splice left-right hand instances sampled from single-hand datasets [43, 70, 14, 71, 65]. It should be noted that although sufficient MoCap data could be collected, considering the cost of the MoCap system, it is still meaningful to utilize single-hand data. Fig. 5 (a) visualizes the distribution of related two hand datasets [43, 57] and *Two-hand 500K* , showing that our proposed dataset is more diverse than them.

Less penetration. For the marker-based data, we fit MANO parameters [49] from 3D skeleton by solving inverse kinematics (IK). As the fingers are often tangled together, penetration and dysmorphism always exist between two hands. To ensure physical interaction, we use the physics engine [1] to optimize the fitted hand pose. Similar to [68, 20], we adopt a sampling-based optimization scheme to iteratively refine interaction. The same strategy is also applied to the splicing process to ensure the assemblies are plausible. Fig. 5 (b) demonstrates the interaction before and after optimization. For more details about our dataset, please refer to Sup. Mat .

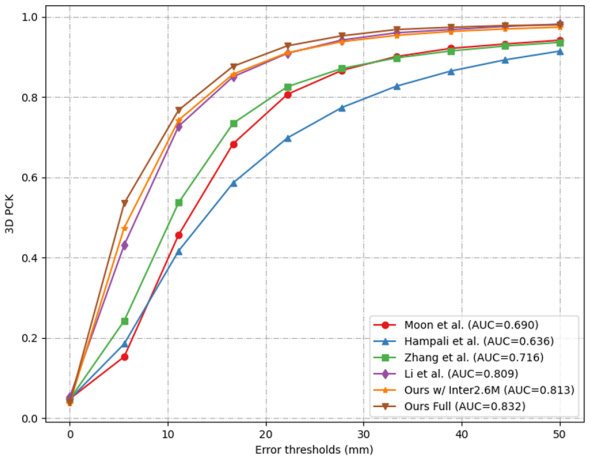


Figure 6. **Comparisons with the SOTA methods on *InterHand2.6M* dataset**. We construct interaction prior using *InterHand2.6M* for fair comparisons. Among them, Moon *et al*. refer to [43], Hampali *et al*. refer to [16], Zhang *et al*. refer to [64] and Li *et al*. refer to [34]

Methods.	MPJPE(mm)↓	MPVPE(mm)↓
Zimmermann *et al*. [70]	36.36	-.
Zhou *et al*. [69]	23.48	23.89
Boukhayma *et al*. [5]	16.93	17.98
Spurr *et al*. [53]	15.40	-
Moon *et al*. [43]	16.53	-.
Hampali *et al*. [16]	20.54	22.27
Fan *et al*. [12]	15.37	-.
Zhang *et al*. [64]	13.77	14.40
Kim *et al*. [27]	12.52	-.
Li *et al*. [34]	8.67	8.92
Ours w/ *Inter2.6M* [43]	8.77	8.81
Ours	**8.34**	**8.51**

Table 1. **Quantitative comparisons on [43]**. Row1-Row4 report pose accuracy with single-hand methods. These results are taken from [34, 64]. The following rows report two-hand methods. Among them, we use MANO joint angles representation to evaluate [16].

4. Experiments

4.1. Datasets and Metrics

Prior data. Data from three different domains are used to construct interaction prior in our practice: marker-based data (*Two-hand 500K*), marker-less data [43] and hands-object data [16, 32]. Since the relative translation is estimated by the network, all two-hand data in [16, 32] can be used to construct interaction prior, even if the hands are not strictly interacting. We use the physics engine to process implausible interactions in the dataset.

Reconstruction data. We only use [43] to train the procedure from image to reconstruction. Before training, we pick out interacting instances and corresponding labels annotated by human and machine (H+M), containing 366K training instances and 261K testing instances. Interacting subjects in [57] are only employed to demonstrate qualitative performance.

Metrics. To evaluate the accuracy of hand pose, we use the mean per joint position error (*MPJPE*) in millimeters. For fair comparisons, we calculate the *MPJPE* after aligning the root joints and scaling the bone lengths of each hand [64, 34, 16, 43]. Apart from that, the percentage of correct keypoints (*PCK*) and the area under the curve (*AUC*) in the range of 0 to 50 millimeters are taken to assess the evaluations. We also report the mean per vertex position error (*MPVPE*) to evaluate the quality of the reconstructed hand surface.

4.2. Comparison with the SOTA

We use the hand pose representation of MANO parameters to report the comparison results. In fairness, the following comparisons are performed on the basis of constructing interaction prior with only the (ALL) branch of [43].

Qualitative results. Comprehensive comparisons show the satisfactory performance of our method. Fig. 7 demonstrates the qualitative results compared with previous SOTA methods [16, 64, 34] for interacting hands reconstruction. Compared to them, our reconstructed interaction generates less penetration while ensuring fidelity, which means that the constructed interaction prior effectively addresses the problem caused by severe occlusion and homogeneous appearance. Fig. 8 further demonstrates the qualitative analysis of the reconstruction. For the completely occluded instances in the gray background, the reconstructed interaction is also consistent with the ground truth (shown on the top right). More results on [43] and [57] can be seen in Sup. Mat .

Quantitative results. The summarized results shown in Tab. 1 indicate the superior performance of our method. From the first four rows in Tab. 1, we present the reconstruction performance of single-hand reconstruction methods [70, 69, 5, 53]. The poor performance suggests that applying the single-hand reconstruction method directly to our task is undesirable due to heavy self-occlusion and appearance confusion. We further compare with almost all recent two-hand reconstruction methods [43, 12, 64, 27, 34, 16] in the community. In exception to them, [39] is not considered because they actually estimate single hand by interacting hand de-occlusion and removal. Compared to our full model, the interaction prior corresponding to w/*Inter2.6M* is only trained with [43]. It can be concluded that non-image paired multimodal training data is positive for constructing interaction prior. Fig. 6 depicts the *PCK* curve of our method, which is also superior to other methods.

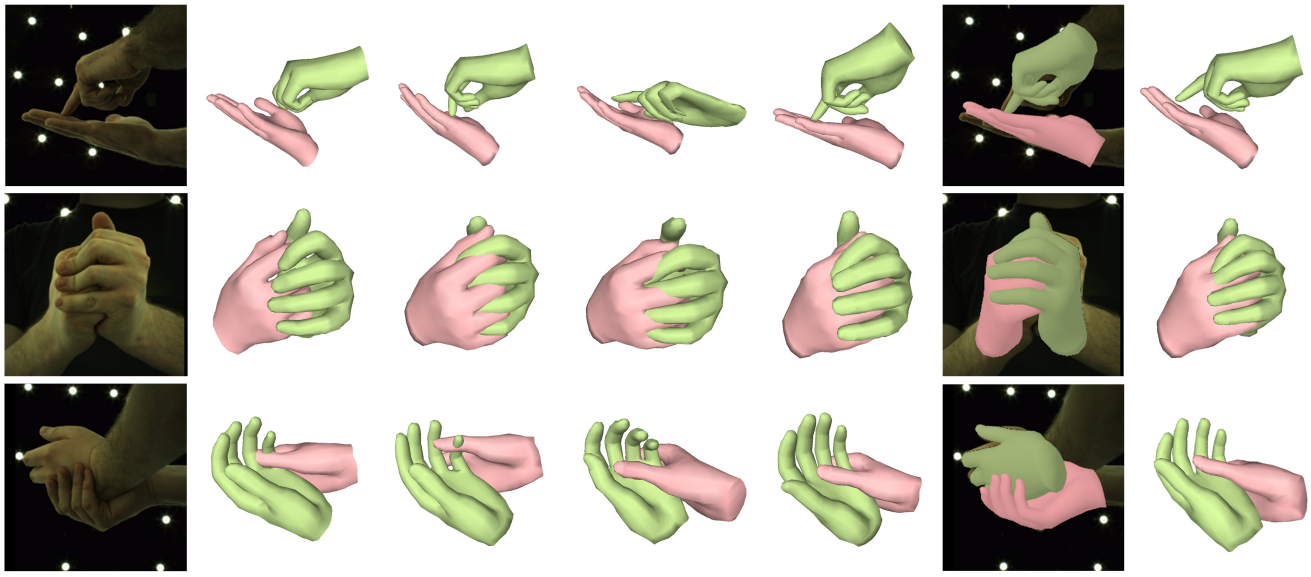


Figure 7. **Qualitative results on [43]**. Qualitative comparisons with other SOTA methods, including Hampali *et al*. [16], Zhang *et al*. [64] and Li *et al*. [34], these comparisons demonstrate our method gains more accurate and high-fidelity reconstruction results.

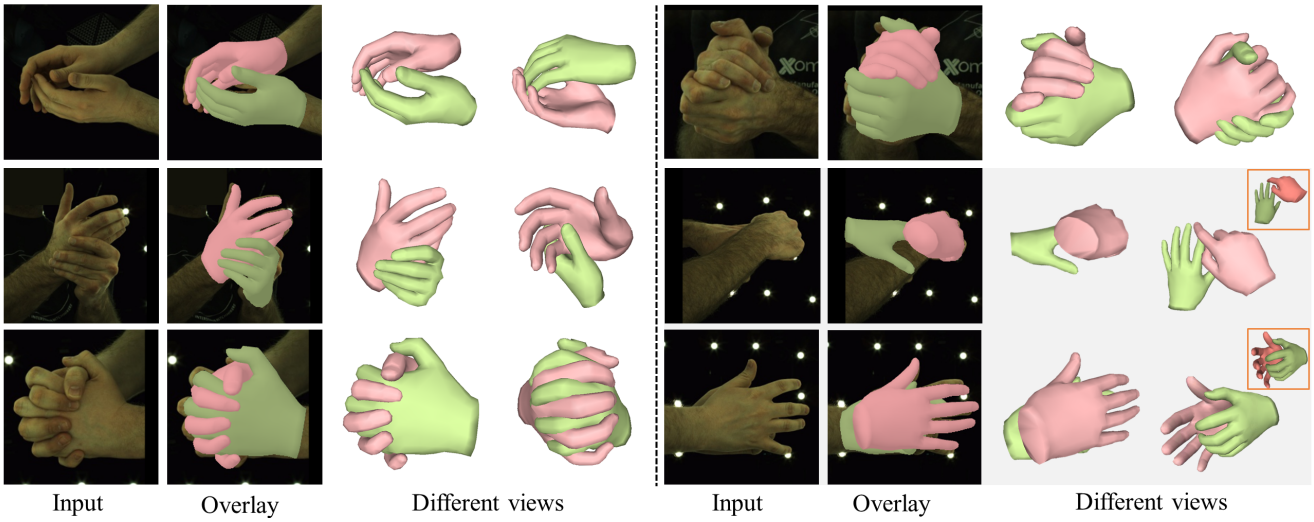
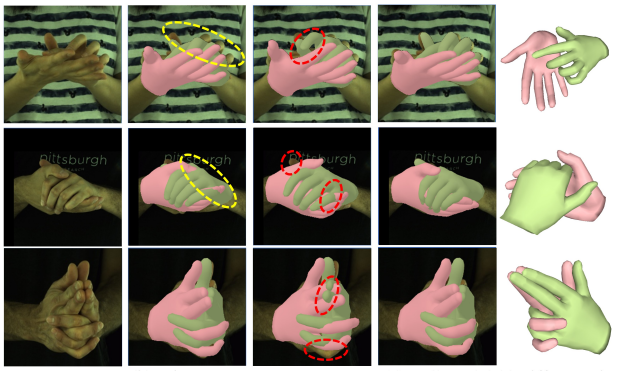


Figure 8. **More reconstruction results on [43]**. High-quality reconstruction results certify the effectiveness of our method. Benefiting from the constructed interaction prior, even with complete self-occlusion (shown in the last two instances), the performance is still satisfactory. The related ground truth is depicted on the top right.

4.3. Ablation Study

We use the hand pose representation of MANO parameters to ablate the effectiveness of each component, but also report the accuracy of other pose representations with the same configuration.

Effectiveness of feature extraction. When extracting interacting features from inputs, we do not overemphasize extracting the unique features for each hand, as the self-occlusion between interacting hands makes it difficult. We design a novel feature extraction module that reflects more global-local context information and report the effectiveness in Tab. 2. We first perform ablation by removing the extracted local features and only using the global features $\mathbf{F}$ to reflect all information. The poor performance in Row.a shows the importance of the feature extraction module, indicating that the extracted features provide more interacting clues. We further ablate the impact of each part in the feature extraction module (Row.b and Row.c) by discarding the corresponding part separately. The visualization about the effectiveness of feature extraction is shown in Fig. 9 (b).



(a) Input (b) w/o Feature (c) w/o Prior (d) Full model (e)Different view

Figure 9. **Effectiveness of different components**. From left to right: The input images, removing feature extraction module, removing constructed interaction prior, reconstructed interaction with our full model, alternative views.

Comp.	Row.	Variants.	MPJPE ↓	MPVPE ↓
2D Feat.	a	w/o feature	10.19	10.25
	b	w/o saliency	9.06	9.11
	c	w/o IAH	9.87	10.07
Hm. Repr.	d	w/ an all-in-one	9.63	9.90
	e	w/ a joint-wise	8.81	9.05
	f	w/ Gaussian	9.22	9.69
Prior	g	w/o prior	11.07	11.49
Fusion	h	w/ ResNet50	9.22	9.27
	i	w/ HRNet32	9.16	9.20
Pose Repr.	j	w/ 3D joints	9.74	-.
	k	w/ 3D vertices	-.	10.52
	l	ours	8.34	8.51

Table 2. **Ablation study of components in our framework**. Adequate ablation experiments have explored the effectiveness of each key component: feature extraction module, IAH, interaction prior, feature fusion network and different representations.

Effectiveness of IAH. To demonstrate that IAH is more suitable for interacting hands reconstruction, we express the heatmaps with different forms and list the corresponding effects. As shown in Row.d and Row.e of Tab. 2, although the conventional heatmaps help to reduce the errors, the impact on reconstruction is more marginal than our proposed IAH. We attribute this success to the adaptability of IAH. Row.f investigates the IAH with Gaussian distribution, and the inferior performance suggests that the Laplacian distribution is more suitable for IAH. More ablations can be found in Sup. Mat .

Effectiveness of prior. Benefiting from the constructed interaction prior, unreal interaction states have been excluded. We analyze the impact of the interaction prior by replacing it with an MLP architecture [51]. Row.g in Tab. 2 shows the result without interaction prior. The unsatisfactory performance highlights the importance of interaction prior, denoting it contributes more to improving accuracy than the feature extraction module. Besides, Fig. 9 (c) further demonstrates the significant decrease in mesh quality.

Effectiveness of ViT-based fusion. The effects of both the extracted features and pre-built prior have been analyzed. Maximizing the performance of extracted features and accurately sampling the constructed prior are critical to the final reconstruction. We compare two other powerful backbones ResNet50 [19] and HRNet32 [54], each of which is applied to completely replace ViT. Comparing Row.h and Row.i in Tab. 2, we see that the ViT-based network gives more powers when fusing interaction features. That is because ViT obtains more global-local context information and effectively models the interactions between two hands. It is noted that only 6 ViT blocks are adopted in the experiments, making the parameter count comparable among different networks and ensuring the fairness of the ablations.

Influence of different representations. We further compare the interaction priors constructed separately with 3D hand joints, 3D hand vertices and MANO parameters within the same dimension. The corresponding performance is reported in Row.j and Row.k of Tab. 2. Among them, the best performance is achieved by the MANO representation, while the lowest accuracy occurs in the representation of 3D vertices. We attribute the reason to the self-restriction of MANO parameters and it is more difficult to embed discrete 3D coordinates.

Discussion on prior structure. We discuss two candidate prior structures: auto-encoder (AE) and variational auto-encoder (VAE). AE is data-dependent and can not generate data. While VAE drives the latent variable to conform to the standard normal distribution and uses reparameterization tricks to improve generativity and robustness [40], which is more reasonable to construct interaction prior.

5. Conclusion

This work treats the interacting hands as a whole, constructs interaction prior based on multimodal datasets, and utilizes joint-wise interaction adjacency to reconstruct interacting hands from monocular images. Compared to most existing works, our framework elegantly combines multimodal datasets to build interaction prior and further recasts the reconstruction as the conditional sampling from this prior. To facilitate its training, *Two-hand 500K* dataset is further constructed with modal diversity and physical plausibility considerations. Our framework based on cross-modal interaction prior would also bring inspiration to other multi-body reconstruction tasks.

Limitations and Future Work. Although we can obtain reasonable interaction from the constructed interaction prior, the penetration is still unavoidable for complicated entanglements. In the future, using the physics engine to guide interaction could bring more benefits to the community.

References

[1] Bullet. https://github.com/bulletphysics/bullet3. 5

[2] Laplace distribution. https://en.wikipedia.org/wiki/Laplace_distribution. 2

[3] Pytorch. https://zh.wikipedia.org/wiki/PyTorch. 4

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 640–653. Springer, 2012. 2

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

[6] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 666–682, 2018. 2

[7] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13274–13283, 2021. 1, 2

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

[9] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, pages 342–359. Springer, 2022. 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*. 2, 4

[11] Zhipeng Fan, Jun Liu, and Yao Wang. Adaptive computationally efficient network for monocular 3d hand pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 127–144. Springer, 2020. 2

[12] Zicong Fan, Adrian Spurr, Muhammed Kocabas, Siyu Tang, Michael J Black, and Otmar Hilliges. Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In *2021 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2021. 1, 2, 6

[13] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2

[14] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019. 5

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2014. 2

[16] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11090–11100, 2022. 1, 2, 6, 7

[17] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, et al. Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Transactions on Graphics (ToG)*, 39(4):87–1, 2020. 2

[18] Rana Hanocka, Gal Metzer, Raja Giryes, and Daniel Cohen-Or. Point2mesh: A self-prior for deformable meshes. *arXiv preprint arXiv:2005.11084*, 2020. 2

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 8

[20] Buzhen Huang, Liang Pan, Yuan Yang, Jingyi Ju, and Yangang Wang. Neural mocon: Neural motion control for physically plausible human motion capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6417–6426, 2022. 5

[21] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Object-occluded human shape and pose estimation with probabilistic latent consistency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):5010–5026, 2022. 2

[22] Buzhen Huang, Tianshu Zhang, and Yangang Wang. Pose2uv: Single-shot multiperson mesh recovery with deep uv prior. *IEEE Transactions on Image Processing*, 31:4679–4692, 2022. 2

[23] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 1, 2, 3

[24] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020. 5

[25] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7122–7131, 2018. 2

[26] Manuel Kaufmann, Emre Aksan, Jie Song, Fabrizio Pece, Remo Ziegler, and Otmar Hilliges. Convolutional autoencoders for human motion infilling. In *2020 International*

Conference on 3D Vision (3DV), pages 918–927. IEEE, 2020. 2

[27] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11189–11198, 2021. 1, 2, 6

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[29] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[30] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5253–5263, 2020. 2

[31] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos 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

[32] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2o: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10138–10148, 2021. 6

[33] Jiaman Li, Ruben Villegas, Duygu Ceylan, Jimei Yang, Zhengfei Kuang, Hao Li, and Yajie Zhao. Task-generic hierarchical human motion prior using vaes. In *2021 International Conference on 3D Vision (3DV)*, pages 771–781. IEEE, 2021. 2

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

[35] 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*, pages 2373–2381, 2021. 2

[36] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1954–1963, 2021. 2

[37] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12939–12948, 2021. 2

[38] Hung Yu Ling, Fabio Zinno, George Cheng, and Michiel Van De Panne. Character controllers using motion vaes. *ACM Transactions on Graphics (TOG)*, 39(4):40–1, 2020. 2, 4

[39] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. *arXiv preprint arXiv:2207.11061*, 2022. 2, 6

[40] Hsien-Yu Meng, Lin Gao, Yu-Kun Lai, and Dinesh Manocha. Vv-net: Voxel vae net with group convolutions for point cloud segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8500–8508, 2019. 8

[41] Paritosh Mittal, Yen-Chi Cheng, Maneesh Singh, and Shubham Tulsiani. Autosdf: Shape priors for 3d completion, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 306–315, 2022. 2

[42] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 5

[43] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 1, 2, 3, 5, 6, 7

[44] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 49–59, 2018. 2

[45] 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 (ToG)*, 38(4):1–13, 2019. 2

[46] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1862–1869. IEEE, 2012. 2

[47] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA 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*, pages 10975–10985, 2019. 2

[48] Mathis Petrovich, Michael J Black, and Gül Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021. 2

[49] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Trans. Gr.*, 36(6):1–17, 2017. 2, 4, 5

[50] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *IEEE International Conference on Computer Vision Workshops*, 2021. 2

[51] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change 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, 2, 5, 8

[52] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 211–228. Springer, 2020. 2

[53] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 89–98, 2018. 2, 4, 6

[54] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 8

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

[56] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. Articulated distance fields for ultra-fast tracking of hands interacting. *ACM Transactions on Graphics (TOG)*, 36(6):1–12, 2017. 2

[57] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 2, 5, 6

[58] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Crossing nets: Combining gans and vaes with a shared latent space for hand pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 680–689, 2017. 2, 4

[59] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dual grid net: Hand mesh vertex regression from single depth maps. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 442–459. Springer, 2020. 2

[60] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Transactions on Graphics (ToG)*, 39(6):1–16, 2020. 2

[61] Linlin Yang, Shile Li, Dongheui Lee, and Angela Yao. Aligning latent spaces for 3d hand pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2335–2343, 2019. 2, 3, 4

[62] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 2, 3, 4

[63] Mingyue Yang, Yuxin Wen, Weikai Chen, Yongwei Chen, and Kui Jia. Deep optimized priors for 3d shape modeling and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3269–3278, 2021. 2

[64] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11354–11363, 2021. 1, 2, 3, 4, 6, 7

[65] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. A hand pose tracking benchmark from stereo matching. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 982–986. IEEE, 2017. 5

[66] Xiong Zhang, Hongsheng Huang, Jianchao Tan, Hongmin Xu, Cheng Yang, Guozhu Peng, Lei Wang, and Ji Liu. Hand image understanding via deep multi-task learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11281–11292, 2021. 2

[67] 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

[68] Zimeng Zhao, Binghui Zuo, Wei Xie, and Yangang Wang. Stability-driven contact reconstruction from monocular color images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1643–1653, 2022. 5

[69] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 1, 2, 6

[70] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 2, 5, 6

[71] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2, 5