RenderIH: A Large-scale Synthetic Dataset for 3D Interacting Hand Pose Estimation

Lijun Li*[1,2], Linrui Tian[1], Xindi Zhang[1], Qi Wang[1], Bang Zhang[1], Liefeng Bo[1], Mengyuan Liu[3], and Chen Chen[4]

[1]Alibaba Group, [2]Shanghai Artificial Intelligence Laboratory, [3]Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University, [4]Center for Research in Computer Vision, University of Central Florida

Abstract

The current interacting hand (IH) datasets are relatively simplistic in terms of background and texture, with hand joints being annotated by a machine annotator, which may result in inaccuracies, and the diversity of pose distribution is limited. However, the variability of background, pose distribution, and texture can greatly influence the generalization ability. Therefore, we present a large-scale synthetic dataset –RenderIH– for interacting hands with accurate and diverse pose annotations. The dataset contains 1M photo-realistic images with varied backgrounds, perspectives, and hand textures. To generate natural and diverse interacting poses, we propose a new pose optimization algorithm. Additionally, for better pose estimation accuracy, we introduce a transformer-based pose estimation network, TransHand, to leverage the correlation between interacting hands and verify the effectiveness of RenderIH in improving results. Our dataset is model-agnostic and can improve more accuracy of any hand pose estimation method in comparison to other real or synthetic datasets. Experiments have shown that pretraining on our synthetic data can significantly decrease the error from 6.76mm to 5.79mm, and our Transhand surpasses contemporary methods. Our dataset and code are available at https://github.com/adwardlee/RenderIH.

1. Introduction

3D interacting hand (IH) pose estimation from a single RGB image is a key task for human action understanding and has many applications, such as human-computer interaction, augmented and virtual reality, and sign language recognition. However, obtaining 3D interacting hand pose annotations from real images is very challenging and time-consuming due to the severe self-occlusion problem. Some previous works [12, 18] have collected some real hand interaction data using a sophisticated multi-view camera system and made manual annotations, but the amount of data is limited. Synthetic 3D annotation data has become increasingly popular among researchers because of its easy acquisition and accurate annotation [27, 22, 3, 7, 15, 24, 40]. However, there remain two main challenges: the validity of the generated 3D hand poses and the diversity and realism of the generated images. Therefore, in this paper, we present a high-fidelity synthetic dataset of 3D hand interaction poses for precise monocular hand pose estimation.

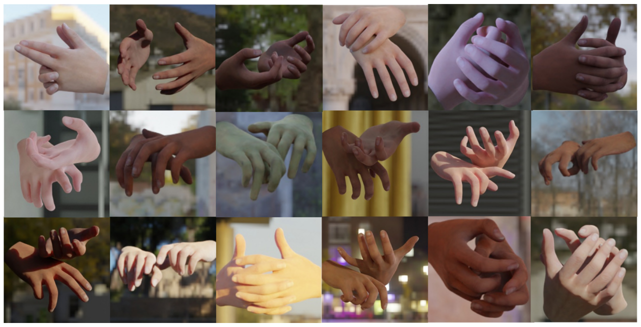

Figure 1. **Randomly selected samples from RenderIH dataset.** The rendered hands are realistic and varied, capturing a variety of poses, textures, backgrounds, and illuminations.

Firstly, ensuring the validity of the generated 3D interacting hand poses is a crucial challenge for a synthetic hand system. For example, the pose of Ego3d [22] is randomized which means a significant portion of the data is not valid. To ensure effective hand interactions, the generated two-hand poses must be proximal to each other, while increasing the risk of hand interpenetration. Therefore, we design an op-

[*]Corresponding author: 4065156@qq.com

Dataset	Type	Data size	MT	AP	background	illumination	Hand type	IH Size
NYU [31]	real	243K	-	✗	lab	uniform	SH	-
STB [39]	real	36K	-	✗	lab	uniform	SH	-
H2O-3D [12]	real	76K	-	✗	lab	uniform	HO	-
H2O [37]	real	571K	-	✗	indoor scenes	uniform	HO	-
MVHM [3]	synthetic	320K	✗	✗	static scenes	uniform	SH	-
ObMan [15]	synthetic	147K	✓	✗	static scenes	uniform	HO	-
DARTset [7]	synthetic	800K	✓	✗	static scenes	manual	SH	-
IH2.6M [26]	real	2.6M	-	✗	lab	uniform	**IH**	628K
Ego3d [22]	synthetic	50K	✗	✗	static scenes	random	**IH**	40K
RenderIH (Ours)	synthetic	1M	✓	✓	HDR scenes	**dynamic**	**IH**	**1M**

Table 1. **Comparison of the related hand datasets.** "MT" is short for multi-textures and means whether the hand models in the dataset are assigned with diverse textures, AP is short for anti-penetration, "Hand type" means which interaction type the dataset focus on (SH-single hand, HO-hand to object, IH-hand to hand), and "IH Size" means the proportion of IH poses. "HDR" is short for High Dynamic Range. Static scenes refer to the use of randomly selected images as the background.

timization process that considers the constraints of hand attraction and anti-penetration in the meantime, to ensure the proximity of two interacting hands and prevent the occurrence of hand penetration (Section 3.1). In addition, the plausibility of hand poses must also be considered. Hence, we introduce anatomic pose constraints and apply adversarial learning to ensure that the generated hand poses adhere to anatomical constraints and realism. Benefiting from pose optimization, our generated dataset contains a rich set of validated two-hand interaction poses as shown in Figure 1.

Secondly, most existing 3D synthetic hand images lack diversity in terms of backgrounds, lighting, and texture conditions, which prevents them from capturing the complex distribution of real hand data [22, 3, 15]. Most existing datasets for hand gesture recognition, such as Ego3d [22], Obman [15], and MVHM [3], do not consider the quality and diversity of the images. For instance, Ego3d [22] uses the same texture as the MANO model [29], which is unrealistic and monotonous. In contrast, our rendering system introduces various textures, backgrounds, and lighting effects that can produce vivid and realistic synthetic hand images (see Section 3.2). By combining HDR background, dynamic lighting, and ray-tracing renderer, we obtain 1M high-quality gesture images (see Figure 1).

To assess the performance of our proposed dataset, we carried out comprehensive experiments on it. We demonstrate how much we can reduce the dependency on real data by using our synthetic dataset. Then we contrast our proposed RenderIH with other 3D hand datasets, such as H2O-3D [12] and Ego3d [22], by training a probing model for each of them and testing on a third-party dataset. Finally, we train a transformer-based network on a mixed dataset of RenderIH and InterHand2.6M (IH2.6M) and achieve state-of-the-art (SOTA) results on 3D interacting hand pose estimation. Our main contributions are as follows:

- We propose an optimization method to generate valid and natural hand-interacting poses that are tightly coupled and avoid interpenetration. For image generation, we design a high-quality image synthesis system that combines rich textures, backgrounds, and lighting, which ensures the diversity and realism of the generated images.
- Based on our data generation system, we construct a large-scale high-fidelity synthetic interacting hand dataset called **RenderIH**, which contains 1 million synthetic images and 100K interacting hand poses. To the best of our knowledge, this is the largest and most high-quality synthetic interacting dataset so far.
- We conduct extensive experiments to verify the effectiveness of our proposed dataset-RenderIH. The results show that with the help of our synthetic dataset, using only 10% of real data can achieve comparable accuracy as the models trained on real hand data. We also propose a transformer-based network that leverages our dataset and achieves SOTA results.

2. Related work

2.1. Realistic hand dataset

Establishing a realistic hand dataset is a tedious and challenging procedure, most realistic data are collected by different sensors [26, 13, 11, 41, 39, 30, 20] including multiple cameras and depth sensors. STB dataset [39] obtained 3D annotations of a single hand (SH) via 2D manual labels and depth data. Since manual annotations are time-consuming [26], some researchers [30, 26, 13, 41] utilized semi-automatic methods to make annotations. Moon et al. [26] captured hand interactions with hundreds of cameras. They manually annotated the 2D keypoints of both hands on a few images and utilized a machine detector to help annotate the rest data. While some researchers [11, 1, 31] proposed automatic methods to make annotations, Hampali et al. [11] collected hand-object (HO) interactions

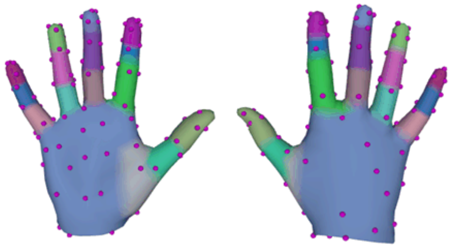

Figure 2. The distribution of anchors and hand subdivision. Purple points denote the anchors.

and jointly optimized 2D key points on multiple RGB-D images to estimate 3D hand poses. Some researchers [8, 37, 9] obtain the 3D annotations of hands via some special equipment. Ye et al. [37] captured hand poses via multiple joint trackers. Due to the limitation of the data collection scene, most realistic datasets are in simple scenarios, e.g. lab [39, 30, 11] or green screen [41, 26, 37, 1, 32]. Most realistic datasets focus on SH or HO interactions and very few papers [26, 32] collect interacting hand data.

2.2. Synthetic hand dataset

To obtain precise annotations and increase the dataset's diversity, several papers [27, 22, 3, 7, 15, 24, 40] established synthetic hand dataset by applying multiple backgrounds [40] or different hand textures [7]. Most datasets [3, 7, 27, 40] focus on SH pose data. DARTset [7] introduced a shaped wrist and rendered hand images with different skins and accessories. But the dataset did not contain IH. To simulate the HO interactions, Hasson et al. [15] utilized physics engine [25] to generate object manipulation poses, but their rendered images are not photo-realistic. Although some datasets [22, 24] provide poses of both hands, the rendered images are not natural enough and lack diversity. Those poses of Ego3d [22] were randomized, which leads to severe interpenetration between hands and the pose is relatively strange. Based entirely on the pose annotations of IH2.6M [26], AIH [24] produced a synthetic interacting hand dataset, but only hand masks were created and other annotations were missing.

We summarize some representative hand datasets and compare them to ours in Table 1. While most datasets focus on SH or HO interactions, they are deficient in handling mesh collision, maintaining high-quality annotations, and providing pose diversity to some extent.

3. RenderIH dataset

One of the main contributions of our paper is the interacting hand pose optimization method that can generate valid and natural poses. In our paper, **valid** poses are non-penetration hands and conform to the anatomic constraint outlined in Table 2. The **natural** poses not only conform to the anatomy but also frequently occur in daily life. We uniformly combine generated poses with a variety of hand textures, high dynamic range (HDR) backgrounds, and camera views. All collections are sampled independently to create images as diverse as possible. In Section 3.1, we introduce our new hand pose generation algorithm. After hand pose generation, how to render the synthetic image is demonstrated in Section 3.2. In Section 3.3, we briefly introduce some statistics about our RenderIH dataset.

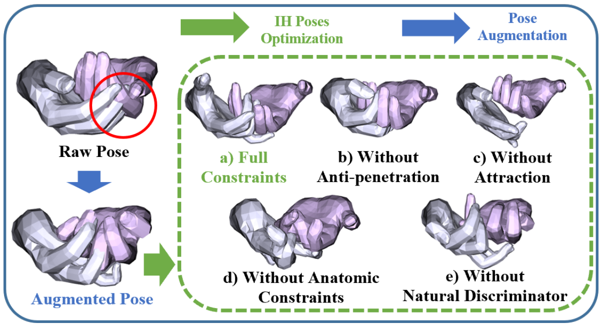

Figure 3. Visualization for the effect of different components in optimization.

3.1. Interacting hand pose optimization

Hand model. Based on the widely used parametric hand model MANO [29], Yang et al. [36] proposed A-MANO, which assigns the twist-splay-bend Cartesian coordinate frame to each joint along the kinematic tree and fit the natural hand better. Therefore, we adopt A-MANO to make our optimization more biologically plausible.

Initial pose generation. To produce massive valid and natural IH interaction poses, we derive raw poses from the IH2.6M [26] and then augment the raw poses by assigning random rotation offsets to hand joints. The augmented poses are shown in Figure 3, after augmentation, the rotation of the j_{th} finger joint can be expressed as:

$$\{R_{ji} \in SO(3)\}_{i=1}^{I} = \{R_j R_b(\theta_i^b) R_s(\theta_i^s)\}_{i=1}^{I}, \quad (1)$$

where I is the number of augmentation, $R_{b/s}(\theta)$ denotes the rotation along the bend/splay axe, the angle offset $\theta^b \in [-90°, 90°]$ and $\theta^s \in [-30°, 30°]$. $SO(3)$ is a group of 3D rotations. $\theta^s = 0$ when the joint is not the finger root joint. To avoid abnormal gestures, each augmented joint is restricted according to Table 2. As the augmented poses are totally random, most of them suffer from serious mesh penetration and their gestures are unnatural, it is necessary to optimize the poses.

Anti-penetration. Inspired by [17], we adopt multi-person interpenetration loss to interacting hands and propose to divide the hand region into 16 parts. Let Ω be the modified Signed Distance Field (SDF) [14] for each hand. The SDF is defined on a voxel grid of dimensions

$N \times N \times N$. It is defined as follows:

$$\Omega(x, y, z) = -min(SDF(x, y, z), 0), \tag{2}$$

where Ω states that its value within a hand is positive and proportional to the distance from the surface, and it is simply zero outside. The penetration loss for a single hand is calculated as follows:

$$L_p^s = \sum_{v \in \{V\}} \Omega_{\hat{s}}(v). \tag{3}$$

V means the hand vertices, s is the side of the hand, and $\hat{s}$ is the side of the other hand. While the hand is highly articulated with a complex pose and shape, basic hand mesh SDF is not accurate enough. We propose to divide the hand into 16 parts based on its joint position and compute a separate Ω function for each hand submesh which is divided according to the hand subdivision in Figure 2. After applying for each submesh, the penetration loss is defined as:

$$L_p^s = \sum_{i=1}^{N} \sum_{j=1}^{N} (\sum_{v \in \{M_{sj}\}} \Omega_{\hat{s}i}(v)), \tag{4}$$

where M_{sj} means the j^{th} submesh of the hand. The total loss of this part is $L_p = L_p^{right} + L_p^{left}$. The detailed visualization comparison between basic SDF loss and our penetration loss is shown in the supplementary material (SM).

Interhand attraction. When the IH is in close contact, severe hand occlusion may occur, making it difficult to make annotations. Additionally, the available close contact data are limited. To address this problem, it is recommended to ensure the IH remains in tight contact.

To create contact between the hands, simply bringing the closest vertices together would suffice. However, to reduce the optimization's time complexity, we adopt anchors to guide the position and pose of both hands. As shown in Figure 2, to downsample the hand vertices as anchors, we traverse IH2.6M to assess the contact frequency of each vertex with the other hand. We selected the vertices with the highest contact frequency as the initial anchors and proceeded to sample the remaining vertices sequentially. Subsequently, we skip the 2-hop neighbors and then continue to sample the yet-to-be-selected ones. Finally, we obtained 108 anchors.

If anchor a_j^l on the left and the anchor a_i^r on the right hand are the closest, they will establish an anchor pair, and the loss of anchor pairs is defined as:

$$L_{ij}^A = \frac{1}{2} k_{ij} {\Delta d_{ij}}^2, \tag{5}$$

where $\Delta d_{ij} = ||a_i^r - a_j^l||_2$. And $k_{ij} = 0.5 * cos(\frac{\pi}{s} \Delta\bar{d}_{ij}) + 0.5$, in which $\Delta\bar{d}_{ij}$ is the initial distance between anchors pair. This definition means the initially close anchors tend to keep in contact. Especially the factor s is set to $0.02m$, and we set $k_{ij} = 0$ if $\Delta\bar{d}_{ij} > s$. The anchor pairs connection and k_{ij} will be rebuilt during the optimization to adapt to dynamically changing IH poses.

finger \ joint	root (B,S)	middle (B)	end (B)
thumb	$[-20, 40]$,$[-30, 30]$	$[-8, 50]$	$[-10, 100]$
index	$[-25, 70]$,$[-25, 15]$	$[-4, 110]$	$[-8, 90]$
middle	$[-25, 80]$,$[-15, 15]$	$[-7, 100]$	$[-8, 90]$
ring	$[-25, 70]$,$[-25, 15]$	$[-10, 100]$	$[-8, 90]$
pinky	$[-22, 70]$,$[-20, 30]$	$[-8, 90]$	$[-8, 90]$

Table 2. **Joint rotation limitations.** The values are in degrees. 'B'/'S' denotes whether the joint can bend/splay.

However, only these constraints cannot keep interacting poses valid with random joint angles, we further introduce anatomic optimization.

Anatomic Optimization. The finger comprises joints, namely the Carpometacarpal joint (CMC), the Metacarpophalangeal joint (MP), and the Interphalangeal joint (IP). According to the coordinates systems of A-MANO, each finger has three joints, and we denote them as root (CMC of thumb, MP of the others), middle (MP of thumb, Proximal IP of the others), and end joint (IP of thumb, Distal IP of the others). Each of them theoretically has 3 DOF. We define the hand pose in Figure 2 as the T-pose, where all rotation angles are zero. The constraints are defined as follows:

- **Available rotation directions.** Middle and end joint can only rotate θ_i^b around the B (Bend) axe, while the root can also rotate θ_i^s around S (Splay) axe. Always keep $\theta_i^t = 0$ around the T (Twist) axe.
- **Angle limitations.** According to hand kinematics [10, 19], the joint rotation limitations are presented in Table 2.

The anatomic optimization objective for each hand is defined as:

$$L_a = \sum_{i=1}^{15} \sum_{a \in \{b,s,t\}} (\beta(\theta_i^a))^2, \tag{6}$$

where $\beta(\theta_i^a) = max(\theta_i^a - \hat{\theta_i^a}, 0) + min(\theta_i^a - \check{\theta_i^a}, 0)$ is the deviation of the rotation angle from its range, and $\hat{\theta_i^a}/\check{\theta_i^a}$ is the max/min value of θ_i^a's range.

Natural discriminator. After anatomic optimization, the poses become valid. However, as shown in Figure 3(e), some optimized poses would not be natural enough. To get the natural poses, we further employ a discriminator $\mathcal{D}$. The detailed structure of the discriminator is illustrated in Figure 4. The single-hand pose Θ is given as input to the multi-layer discriminator. The output layer predicts a value $\in [0, 1]$ which represents the probability of belonging to the natural pose. The objective for $\mathcal{D}$ is:

$$L_{\mathcal{D}} = \mathbb{E}_{\Theta \sim P_R}[(\mathcal{D}(\Theta) - 1)^2] + \mathbb{E}_{\Theta \sim P_G}[\mathcal{D}(\Theta)^2], \tag{7}$$

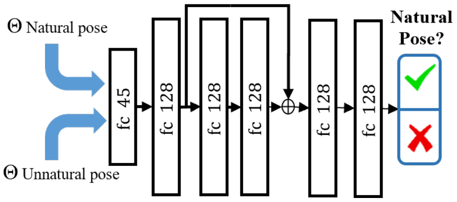

Figure 4. The architecture of the discriminator.

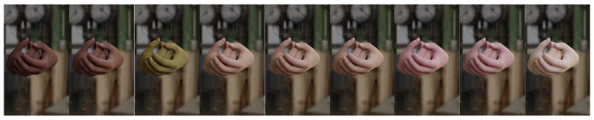

Figure 5. Same hand with different hand textures.

where P_R represents a hand pose from real datasets, such as IH2.6M [26] and Freihand [41], P_G is a generated pose. The adversarial loss that is backpropagated to pose optimization is defined as:

$$L_{adv} = \mathbb{E}_{\Theta \sim P_G}[(\mathcal{D}(\Theta) - 1)^2]. \tag{8}$$

The discriminator is pre-trained before optimization. We extract 63K natural single-hand poses from Freihand [41], DexYCB [2], and IH2.6M [26], their ”natural” probabilities p_n are labeled as 1. To get unnatural poses, we follow the methods in ”Initial pose generation” to randomly add offsets to the poses, and calculate their probabilities according to the offsets (the higher the offsets, the closer the p_n is to 0). The qualitative and quantitive improvements brought by $\mathcal{D}$ could be seen in SM. Since the natural standard may vary from person to person, we also conducted a user study to confirm the discriminator’s effect in SM.

Poses Optimization. In IH optimization, for each hand, it has 15 joints rotation $\Theta = \{R_i \in SO(3)\}_{i=1}^{15}$, hand root rotation $R_r \in SO(3)$ and hand root translation $T_r \in \mathbb{R}^3$, we take $\psi = \{\Theta, T_r\}$ as the optimization parameters and the total IH loss is denoted as:

$$\underset{\psi^r,\psi^l}{argmin}(w_1 \sum_{i=1}^{A_r}\sum_{j=1}^{A_l} L_{ij}^A + w_2 L_a + w_3 L_{adv} + w_4 L_p), \tag{9}$$

where A_r/A_l is the anchor numbers of right/left hand, $L_a = L_a^r + L_a^l$, and w_* is the weight hyperparameter.

3.2. Rendering

Our dataset offers various benefits, including high-resolution hand textures that create a more natural appearance. Additionally, we simulate natural lighting and environments to address limited diversity in studio settings. Furthermore, our dataset covers a wide range of poses and camera positions, bridging the gap between real-world applications and synthetic data.

Texture. To enhance the variety of skin textures we present a broad selection of hues as illustrated in Figure 5. Color tones include white, light-skinned European, dark-skinned European, Mediterranean or olive, yellow, dark brown, and black. A total of 30 textures are available. In addition, random skin tone parameters can be superimposed on these base skin tones in the shaders to adjust brightness, contrast, and more. Apart from that, these textures also depict wrinkles, bones, veins, and hand hairs to cope with differences in gender, ethnicity, and age.

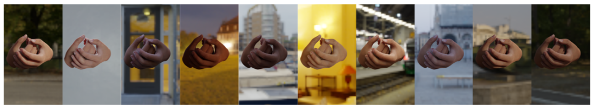

Figure 6. Same hand under diverse illumination.

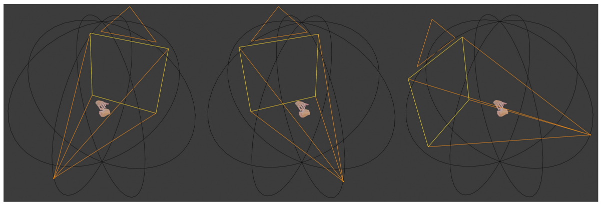

Figure 7. Different viewpoints from the camera track.

Lighting and background. It is widely accepted that high-quality synthetic data should resemble real-world scenes as much as possible. For instance, the authors mixed their synthetic hands images with diverse real-world background photographs when creating IH synthetic data [22]. However, simply pasting the rendered hands on the background images is unnatural due to differences in lighting conditions and light angles. Since creating a large number of various synthetic 3D background models is time-consuming, we composite synthetic hands with various real-world scenery panoramic images. We collect 300 high-dynamic-range (HDR) photography with realistic indoor and outdoor scenes with appropriate lighting for rendering purposes. They enable our hand models to blend seamlessly with diverse settings resulting in highly photorealistic rendered scenes (see Figure 6).

Camera Settings. We defined a spherical camera arrangement that can contain both viewpoints, enhancing the generalization of the model to different viewpoints. The center of the two-handed model is first computed and placed at the center of the world, and the camera track is placed around the center with the camera pointing to the center. Figure 7 shows the layout of our simulation environment. For each pose, we define four 360-degree circular tracks, which can be averaged by the number of samples to define dense or sparse viewpoints. For sparse sampling, 10 viewpoints were selected for each track.

Render quality. Our major objective is to improve the photorealism of the synthetic dataset. Therefore, we render the scene in Blender based on the ray-tracing rendering engine Cycles. When creating the hand mesh, we used cus-

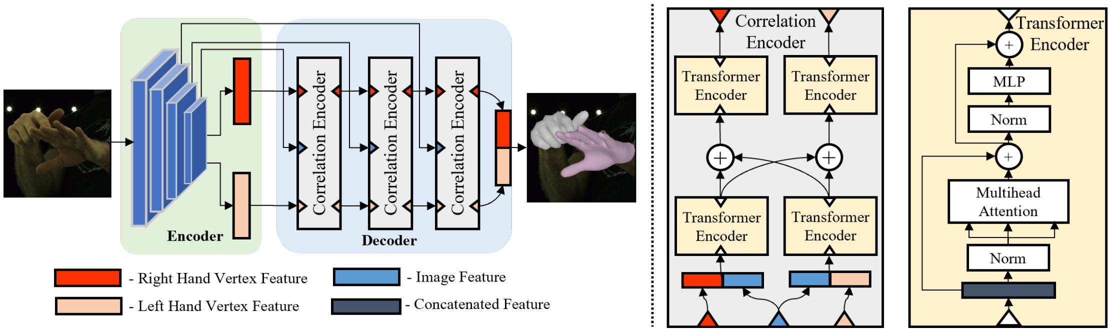

Figure 8. Network architecture. We use the global features extracted by the encoder to predict the left-hand features and right-hand features. After that, our model gradually regresses the hand vertices from 3 identical correlation encoder blocks by fusing multi-resolution image features with hand features. Each correlation encoder contains two transformer encoders and lateral connection from the other hand feature.

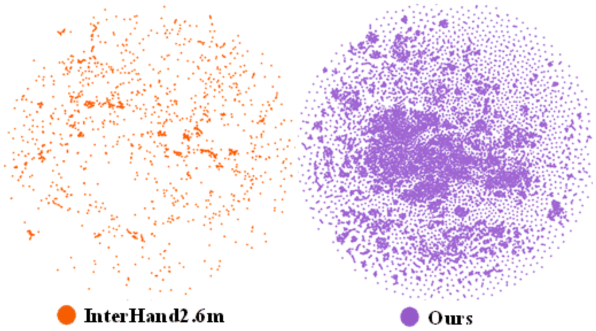

Figure 9. TSNE visualization for IH poses distribution. Our data not only contain the raw poses of IH2.6M but also fill the vacancy by augmentation, resulting in a broader distribution.

tom shader settings to adjust the base color, subsurface, and roughness to make the skin more realistic. The resolution of the image is 512×334 pixels and the color depth is 8 bits.

3.3. Analysis of RenderIH dataset

For distribution diversity comparison, we project the hand pose in IH2.6M and RenderIH into the embedding space using TSNE [34]. Figure 9 clearly shows that our data has a broader pose distribution than IH2.6M. Examples of synthetic images are depicted in Figure 1 and the rendering video can be found in the SM. More visualization effects of different optimization modules and statistical information can be found in the SM.

4. TransHand

We propose a transformer-based network, TransHand, for 3D interacting hand pose estimation and conduct extensive experiments on it.

As the transformer blocks are effective in modeling global interactions among mesh vertices and body joints [23, 35], we propose a transformer-based IH network. Our system contains two parts: the encoder and the decoder. Given an image with size 256×256, the encoder outputs a global feature vector G_F and the intermediate feature maps $\{F_i, i = 1, 2, 3\}$ where i indicates the feature level. After that, we map G_F to the left vertex feature L_F and the right vertex feature R_F by using fully connected layers. Since the global feature does not contain fine-grained local details, we concatenate different level features F_i with the hand vertex feature as input to the decoder blocks.

As shown in Figure 8, the decoder consists of 3 identical blocks. Each block consists of 2 sub-modules, each sub-module is a typical transformer encoder composed of a multi-head attention module and an MLP layer. Each block is made up of two transformer encoders. As there is usually mutual occlusion in IH, it is natural to combine the other hand feature to improve the estimation precision. Inspired by Slowfast [6], we use a symmetric structure to incorporate the other hand feature by adding it, which is the lateral connection in the Correlation Encoder (CE) shown in Figure 8. Each block has three inputs including the left vertex feature, right vertex feature, and image feature. The blocks gradually upsample the coarse mesh up to refined mesh and finally to the original dimension with 778 vertices.

Loss Function. For training, we apply L_1 loss to 3D mesh vertices and hand joints, and L_1 loss to 2D projected vertices and hand joints.

$$L_{joint} = \sum_{s=0}^{1} \sum_{i=0}^{M-1} \sum_{d \in \{3D,2D\}} \| J^{d}_{s,i} - J^{d,GT}_{s,i} \|_1, \tag{10}$$

$$L_{mesh} = \sum_{s=0}^{1} \sum_{i=0}^{N-1} \sum_{d \in \{3D,2D\}} \| V^{d}_{s,i} - V^{d,GT}_{s,i} \|_1, \tag{11}$$

where s represents the hand side, i represents the number of joints or vertices, and d denotes whether the computation is for 3D or 2D. To guarantee the geometric continuity of the predicted vertices, smoothness loss is applied which regularizes the consistency of the normal direction between the

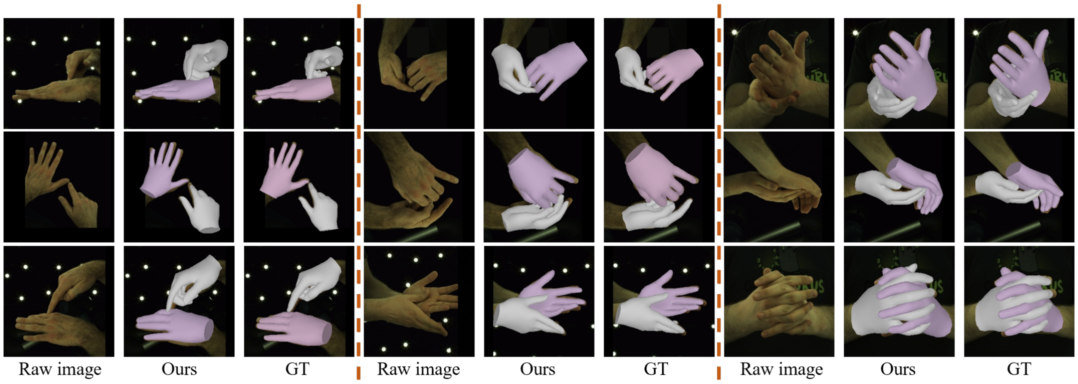

Figure 10. Qualitative results of our method on IH2.6M test set.

predicted and the ground truth mesh:

$$L_{smooth} = \sum_{s=0}^{1} \sum_{f=0}^{F-1} \sum_{j=0}^{2} \|e_{f,j,s} \cdot n_{f,s}^{GT}\|_1, \tag{12}$$

where f means the face index of hand mesh, j means the edge of face f and $n^{GT}$is the GT normal vector of this face.

5. Experiments

5.1. Experiment setup

Dataset. IH2.6M [26] is the largest real dataset with interacting hand (IH), and most of our experiments are conducted on this dataset. As we are only focused on IH, we selected only the IH data with both human and machine annotations. After discarding single-hand samples and invalid labeling, we obtain 366K training samples and 261K testing samples. Tzionas dataset [33] is a small IH dataset. We only use it for generalization ability evaluation by using the models trained from different datasets. H2O-3D [12] is a real dataset with 3D pose annotations for two hands and an object during interactions. It contains 60K samples. Ego3d [22] provides 50K synthetic images and corresponding labels of two hands, in which 40K samples are IH and the poses are randomized.

Implementation details. The input images are resized to 256 × 256 and fed to TransHand encoder to generate the global feature and image feature maps. ResNet50 [16] is selected as the encoder. For all experiments, the networks are implemented using Pytorch [28]. We train all models with IH images using Adam optimizer. The initial learning rate is $1e^{-4}$ and the batch size is 64. All experiments are performed on 1 NVIDIA Ampere A100 GPU. To demonstrate the usefulness of our RenderIH, we train three mainstream IH pose estimation methods on IH2.6M and a combination of IH2.6M and RenderIH, InterNet[1] [26], DIGIT[1] [4] and state-of-the-art method IntagHand[2] [21].

With $\mathcal{D}$	No $\mathcal{D}$	Raw poses	Augmented poses
81.25%	54.68%	90.82%	32.92%

Table 3. User study on natural rate. The higher the number, the more natural it is.

Evaluation metrics. To evaluate these methods, we report results by two standard metrics: Mean Per Joint Position Error (MPJPE) and the Mean Per Joint Position Error with Procrustes Alignment (PAMPJPE) in millimeters (mm). Additionally, to ensure a fair evaluation with prior research [21, 38], we select the MCP joint of middle finger as root joint and also report SMPJPE which performs scaling to the ground truth bone length. To evaluate the accuracy of estimating the relative position between the left and right hand roots during interaction, we utilize the mean relative-root position error (MRRPE) [4] and hand-to-hand contact deviation (CDev) [5] metrics. More results are presented in SM with wrist as root joint for future comparison.

5.2. Results and analysis

User study for naturalness. Since the perceptions of ”natural” may differ from human to human, We conduct experiments to prove the discriminator's effect. We invited 20 persons with/without computer technical background, their ages are from 20 to 60, and the proportion of male to female was approximately 2:1. For each of them, we show 120 pictures (including 30 of augmented poses, 30 of optimized poses, 30 of optimized without discriminator, and 30 of raw poses from IH2.6M) of the IH poses, they are asked to determine whether the shown poses are natural, we count the NR (natural rate) of each category. The results are presented in Table 3, the “Raw poses” are those from IH2.6M[26], they are performed by humans and have high NR, however, some serious mesh-penetration caused by annotation mistakes might make the testers hardly to determine the “natural”. The “Augmented poses” are augmented from the raw

[1] Since InterNet and DIGIT are trained on the IH subset of IH2.6M v0.0, we train them on v1.0 to make fair comparisons.

[2] All the training codes have been open-sourced by the authors.

poses by assigning random rotation offsets to hand joints, they follow the joint limitation but have randomness, and some of them are in mesh penetration, the NR is low in this category. Optimizing the augmented poses without $\mathcal{D}$ solves the penetration, and the poses are valid, but the poses are not natural enough. It is clear that $\mathcal{D}$ improves the naturalness of the poses.

Effectiveness of correlation encoder. Table 4 shows the performance of models with and without the CE. The baseline method fuses the left-hand feature and right-hand feature with the image feature independently through a transform encoder. The result indicates CE can improve performance by fusing the correlation between hands. Our model is used as default model for subsequent experiments.

method \ metric	PAMPJPE/MPJPE/SMPJPE(mm)↓
Baseline	7.32/11.12/10.82
Baseline+CE	6.76/10.6/9.63

Table 4. Effect of correlation encoder (CE) on IH2.6M test set (PAMPJPE/MPJPE/SMPJPE(mm)↓). It is shown that CE helps reduce the error by a clear margin.

method \ trainset	IH2.6M	Mixed
InterNet	18.28	17.19
DIGIT	15.48	14.28
IntagHand	10.9	9.72
Ours	10.6	10.06

Table 5. Comparison between models trained from IH2.6M and a mixture of RenderIH and IH2.6M in MPJPE(mm)↓. The methods are reproduced using their official training code.

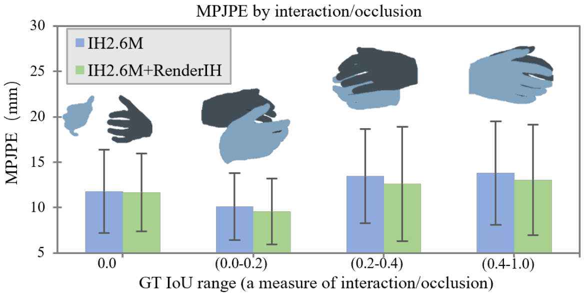

Figure 11. Comparing MPJPE by the degree of occlusion on Tzionas dataset. The IoU between groundtruth left/right masks measures the degree of interaction. The left (yellow) and right (blue) hand masks provide interaction examples in each IoU range

Mixing synthetic with real images. To demonstrate the usefulness of RenderIH, we test InterNet, DIGIT, IntagHand, and our TransHand on the IH2.6M test set under the setting of training with or without using the full 1M data from the RenderIH dataset. As shown in Table 5, RenderIH is helpful to further reduce the estimation error. For example, the error can be greatly reduced from 10.9mm to 9.72mm for the SOTA IntagHand method. The results prove that our RenderIH has great complementarity with real data. Meanwhile, when hand-hand occlusion is severe, training with our synthetic dataset can handle those cases better than IH2.6M only which is shown in Figure 11.To quantify the impact of interaction and occlusion, we use the IoU between left and right hand ground truth masks following DIGIT [4]. The higher IoU implies more occlusion and half-length of the error bars correspond to 0.5 times of MPJPE standard deviation. With minimal occlusion, the MPJPE is similar between the mixed image model and IH2.6M only. As occlusion increases, the mixed image model reduces MPJPE more substantially than IH2.6M alone. This highlights the value of our RenderIH data.by ambiguities.

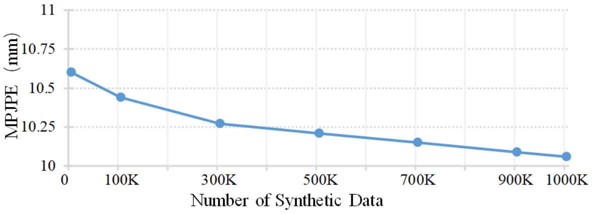

Figure 12. Results of training IH2.6M with different number of RenderIH images on MPJPE(mm)↓.

Synthetic data size influence. During the training phase that involved various combinations of synthetic data and the IH2.6M training set, an obvious decline in the error is observed initially, followed by a gradual decrease after the incorporation of 900K synthetic images, as illustrated in Figure 12. The trend indicates that beyond a certain volume of synthetic data, the benefits of incorporating additional data become marginal. To balance the cost of training and accuracy, we select 1M as the optimal size for RenderIH.

Training strategy comparison. The training strategy of synthetic data and real data is studied in this section. From Figure 13, both data mix training and pretraining from synthetic data can lead to significantly higher accuracy. Compared to dataset mixing, pretraining on the synthetic followed by fine-tuning on real images led to better precision. In contrast to dataset mixing, our results suggest that pretraining on synthetic data followed by finetuning on real images offers a more effective approach for reducing error.

Real data size influence. We study how the real data size affects the estimation precision in Figure 13. We use all the samples from RenderIH in this section. For real data, we sample the number of data ranging from 3663 to 366358, which takes 1%, 5%, 10%, 30%, 50%, 70%, and 100% of the real data. Although training only on RenderIH performs poorly, the MPJPE can be greatly reduced from 27.73mm to 12.6mm by finetuning on only 1% of real data. With finetuning on 10% of real data, the MPJPE can be almost the same as training on the full real data. When finetuning on all real data, the error can be 0.96mm lower than training

Method	PAMPJPE↓	MPJPE↓	SMPJPE↓	MRRPE↓	CDev↓
InterNet* [26]	11.72	18.28	16.68	-	-
DIGIT* [4]	9.72	15.48	13.43	-	-
InterShape [38]	-	-	13.07	-	-
HDR [24]	-	13.12	-	-	-
IntagHand [21]	6.10	10.30	8.79	12.1	25.1
IntagHand*	7.16	10.90	10.47	13.6	29.6
Ours	6.76	10.66	9.63	12.98	27.9
Ours$^{\#}$	5.79	9.64	8.18	11.95	24.6

Table 6. Comparing with SOTA methods on IH2.6M test set ($*$ means official code reproduction, $\#$ means RenderIH pretraining)

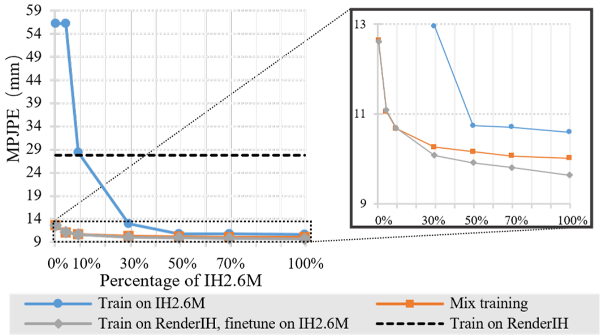

Figure 13. Comparison between training with RenderIH only, with part of IH2.6M only, the combination of the two, with pretraining on RenderIH and finetuning on IH2.6M.

trainset \ testset	IH2.6M	Tzionas
H2O-3D	47.81/48.19	35.61/34.01
Ego3d	58.40/57.48	56.90/54.89
RenderIH	43.35/42.17	32.80/28.70

Table 7. Generalization ability comparison between H2O-3D, Ego3d, and RenderIH on MPJPE/SMPJPE (mm)↓. The number of samples is 40K and fixed for each dataset.

trainset \ testset	IH2.6M	Tzionas
H2O-3D+IH2.6M	11.05/9.91	12.03/12.02
Ego3d+IH2.6M	10.66/9.60	11.13/11.06
RenderIH+IH2.6M	10.58/9.52	10.63/10.56

Table 8. Training on the mixture of datasets with all IH2.6M data on MPJPE/SMPJPE (mm)↓. The number of samples is 40K for each dataset.

only on all real data.

Comparison with H2O-3D dataset, Ego3d dataset and RenderIH subset. In Table 7 and Table 8, we compare the generalization ability of these datasets with the same number of 40K samples. The model pretrained on RenderIH reaches lower error than other models pretrained on H2O-3D and Ego3d in Table 7, which proves that our artificial data is realistic and the knowledge is more easily transferable. The model trained on RenderIH performs better, possibly because all images have objects that interfere with two-handed interaction in H2O-3D. When training TransHand on RenderIH and IH2.6M, the estimation error is the lowest both in the IH2.6M and Tzionas dataset which is shown in Table 8. Especially the result on Tzionas dataset shows our varied pose distribution, background, and texture is helpful for improving generalization.

Comparison with SOTA methods. As is shown in Table 6, our TransHand can outperform SOTA IntagHand method trained from its official code. Furthermore, their method involves multitask learning and their network comprises of complex graph transformer modules. In comparison, our method is simpler yet highly effective. When pretraining on RenderIH and finetuning on the IH2.6M data, our method can further reduce the MPJPE by about 1mm. Better hand-hand contact (CDev) and better relative root translation (MRRPE) can be observed in this table. Moreover, it is shown in Table 9 that training on our dataset in addition to IH2.6M can lead to obviously lower error on the Tzionas dataset compared with training on IH2.6M alone.

Metrics	MPJPE/MRRPE/CDev↓
Training set \ Test set	Tzionas
IH2.6M	11.38/11.1/19.9
IH2.6M+RenderIH	10.49/9.37/19.5

Table 9. The comparison of training with or without our dataset on Tzionas dataset.

Qualitative results. Our qualitative results are shown in Figure 10. We can see our method can generate high-quality IH results in IH2.6M images. More in-the-wild results can be found in the SM.

6. Conclusion

In this paper, we propose a new large-scale synthetic dataset for 3D IH pose estimation. Various experiments are conducted to study the effectiveness of RenderIH. With the whole synthetic hand images and only 10% of real hand images, we can achieve precision that is comparable to the same method which is trained on all the real hand images. We hope that this dataset could be a meaningful step towards developing 3D IH pose estimation models that do not depend on real data and adaptable to to various scenes.

References

[1] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 361–378, Cham, 2020. Springer International Publishing. 2, 3

[2] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021. 5

[3] Liangjian Chen, Shih-Yao Lin, Yusheng Xie, Yen-Yu Lin, and Xiaohui Xie. Mvhm: A large-scale multi-view hand mesh benchmark for accurate 3d hand pose estimation. pages 836–845, 01 2021. 1, 2, 3

[4] 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 *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*, pages 1–10. IEEE, 2021. 7, 8, 9

[5] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 7

[6] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. 6

[7] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. Dart: Articulated hand model with diverse accessories and rich textures, 10 2022. 1, 2, 3

[8] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 3

[9] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Robust hand pose regression using convolutional neural networks. In Anibal Ollero, Alberto Sanfeliu, Luis Montano, Nuno Lau, and Carlos Cardeira, editors, *ROBOT 2017: Third Iberian Robotics Conference*, pages 591–602, Cham, 2018. Springer International Publishing. 3

[10] Verónica Gracia-Ibáñez, Margarita Vergara, Joaquín L Sancho-Bru, Marta C Mora, and Catalina Piqueras. Functional range of motion of the hand joints in activities of the international classification of functioning, disability and health. *Journal of Hand Therapy*, 30(3):337–347, 2017. 4

[11] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 2, 3

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

[13] Shangchen Han, Beibei Liu, Randi Cabezas, Christopher D. Twigg, Peizhao Zhang, Jeff Petkau, Tsz-Ho Yu, Chun-Jung Tai, Muzaffer Akbay, Zheng Wang, Asaf Nitzan, Gang Dong, Yuting Ye, Lingling Tao, Chengde Wan, and Robert Wang. Megatrack: Monochrome egocentric articulated hand-tracking for virtual reality. *ACM Trans. Graph.*, 39(4), aug 2020. 2

[14] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J Black. Resolving 3d human pose ambiguities with 3d scene constraints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2282–2292, 2019. 3

[15] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 04 2019. 1, 2, 3

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 7

[17] 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. 3

[18] Hanbyul Joo, Tomas Simon, Mina Cikara, and Yaser Sheikh. Towards social artificial intelligence: Nonverbal social signal prediction in a triadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10873–10883, 2019. 1

[19] Derek G Kamper, T George Hornby, and William Z Rymer. Extrinsic flexor muscles generate concurrent flexion of all three finger joints. *Journal of biomechanics*, 35(12):1581–1589, 2002. 4

[20] 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 (ICCV)*, pages 10138–10148, October 2021. 2

[21] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2761–2770, 2022. 7, 9

[22] 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. 1, 2, 3, 5, 7

[23] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 12919–12928. IEEE, 2021. 6

[24] 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. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 380–397, Cham, 2022. Springer Nature Switzerland. 1, 3, 9

[25] Andrew Miller and Peter Allen. Graspit!: A versatile simulator for robotic grasping. *Robotics Automation Magazine, IEEE*, 11:110 – 122, 01 2005. 3

[26] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision*, pages 548–564. Springer, 2020. 2, 3, 5, 7, 9

[27] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 1163–1172, Oct 2017. 1, 3

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035, 2019. 7

[29] Javier Romero, Dimitrios Tzionas, and Michael Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36, 11 2017. 2, 3

[30] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4645–4653, 2017. 2, 3

[31] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. volume 33, 08 2014. 2

[32] 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. 3

[33] 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*, 118:172–193, 2016. 7

[34] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 6

[35] Ziniu Wan, Zhengjia Li, Maoqing Tian, Jianbo Liu, Shuai Yi, and Hongsheng Li. Encoder-decoder with multi-level attention for 3d human shape and pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13033–13042, 2021. 6

[36] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. pages 11077–11086, 10 2021. 3

[37] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 2, 3

[38] Baowen Zhang, Yangang Wang, Xiaoming Deng, Yinda Zhang, Ping Tan, Cuixia Ma, and Hongan Wang. Interacting two-hand 3d pose and shape reconstruction from single color image. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 11334–11343. IEEE, 2021. 7, 9

[39] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016. 2, 3

[40] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4913–4921, 2017. 1, 3

[41] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max J. Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 2, 3, 5