

3D Interacting Hand Pose Estimation by Hand De-occlusion and Removal

Hao Meng^{1,3*}, Sheng Jin^{2,3*}, Wentao Liu^{3,4}, Chen Qian³
 Mengxiang Lin¹, Wanli Ouyang^{4,5}, and Ping Luo²

¹ Beihang University ² The University of Hong Kong ³ SenseTime
 Research and Tetras.AI ⁴ Shanghai AI Lab ⁵ The University of Sydney
 hao_meng@163.com, {jinsheng, liuwentao, qianchen}@tetras.ai
 linmx@buaa.edu.cn, wanli.ouyang@sydney.edu.au, pluo@cs.hku.hk

Abstract. Estimating 3D interacting hand pose from a single RGB image is essential for understanding human actions. Unlike most previous works that directly predict the 3D poses of two interacting hands simultaneously, we propose to decompose the challenging interacting hand pose estimation task and estimate the pose of each hand separately. In this way, it is straightforward to take advantage of the latest research progress on the single-hand pose estimation system. However, hand pose estimation in interacting scenarios is very challenging, due to (1) severe hand-hand occlusion and (2) ambiguity caused by the homogeneous appearance of hands. To tackle these two challenges, we propose a novel Hand De-occlusion and Removal (HDR) framework to perform hand de-occlusion and distractor removal. We also propose the first large-scale synthetic amodal hand dataset, termed Amodal InterHand Dataset (AIH), to facilitate model training and promote the development of the related research. Experiments show that the proposed method significantly outperforms previous state-of-the-art interacting hand pose estimation approaches. Codes and data are available at <https://github.com/MengHao666/HDR>.

Keywords: 3D Interacting Hand Pose Estimation, De-occlusion, Removal, Amodal InterHand Dataset

1 Introduction

Estimating the 3D hand pose from a monocular RGB image is critical in many real-world applications, *e.g.* human-computer interaction, augmented and virtual reality (AR/VR), and sign language recognition. Although significant progress has been made for single-hand pose estimation, analysis of hand-hand interactions remains challenging. Estimating 3D interacting hand pose from a single RGB image has attracted increasing research attention in recent years.

In this paper, we propose to decompose the challenging interacting hand pose estimation task, and predict the pose of the left and the right hand separately.

* Equal Contribution.

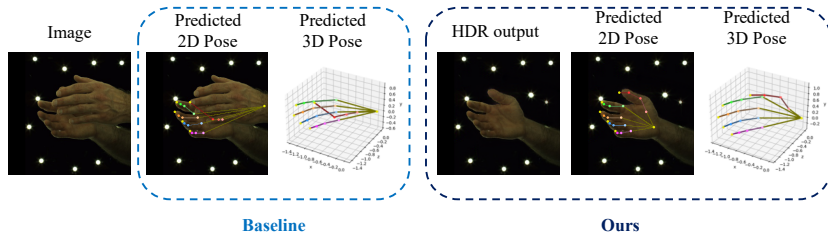


Fig. 1. State-of-the-art hand pose estimation models often struggle to estimate the 3D poses of interacting hands, due to severe hand-hand occlusion and appearance ambiguity of two hands. In this example, we observe erroneous pose estimation of the occluded part and significant uncertainty between the left and right wrist. Our HDR framework tackles these two challenges via hand de-occlusion (recovering the appearance content of the occluded part) and removal (removing the other distracting hand). It transforms the challenging interacting hand image into a simple single-hand image, which can be easily handled by the hand pose estimator.

However, solving single-hand pose estimation in close two-hand interaction cases is non-trivial, because of two major challenges. One of the main challenges is the severe hand-hand occlusion. Considering hands in close interactions, the occlusion patterns are complex. Many areas of the target hand can be occluded, making it very challenging to infer the pose of the invisible parts. Another challenge is that the homogeneous and self-similar appearance of hands (*i.e.* the left and the right hand) may cause ambiguities. And the hand pose estimator may be confused by the other visually similar distracting hand.

To tackle these challenges, we propose a simple yet effective Hand De-occlusion and Removal (HDR) framework. Specifically, our HDR framework comprises three parts, Hand Amodal Segmentation Module (HASM), Hand De-occlusion and Removal Module (HDRM), and the Single Hand Pose Estimator (SHPE). HAS segments both the complete (amodal) and visible parts for both two hands. The resulting segmentation masks not only contain information to localize the rough position of the two hands, but also provide cues for subsequent de-occlusion and removal process by HDRM. *De-occlusion* targets at predicting the appearance content of the occluded part. *Removal* targets at removing the distracting part in the image. In our case, when estimating the pose of the right hand, the left hand becomes the distracting part and should be removed. As shown in Fig. 1, recent state-of-the-art hand pose estimation methods suffer from severe hand-hand occlusion and the homogeneous appearance of two hands, resulting in inferior pose estimation results. Thanks to our proposed HDR framework, we can transform the challenging scenario of hand-hand interactions into a common single-hand scenario, which can be easily handled by an off-the-shelf SHPE.

However, to the best of our knowledge, there exist no datasets that contain both the amodal segmentation and appearance content ground-truths of interactive hands. To fill in this blank, we synthetically generate a large-scale Amodal

InterHand dataset, namely AIH dataset. The dataset contains over 3 million interacting hand images along with ground-truth amodal and modal segmentation, de-occlusion and removal ground-truths. The dataset consists of two parts, *i.e.* AIH_Syn and AIH_Render. AIH_Syn is obtained by simple random copy-and-paste. It retains detailed and realistic appearance information. However, it may generate implausible interacting hand poses that violate the biomechanical structure of the human body. AIH_Render is generated by rendering the textured 3D interacting hand mesh to the image plane. The inter-dependencies between two hands are fully considered to avoid physically implausible configurations, *e.g.* intersecting fingers. However, it may suffer from the appearance gap because the rendered texture is synthetic. Combining the advantages of both, we make a large-scale 3D hand-hand interaction dataset with large pose and appearance variety. We empirically validate the effectiveness of the synthetic dataset through extensive experiments. We envision that our proposed dataset will foster the development of the related research, *e.g.* interacting hand pose estimation, amodal or modal instance segmentation, de-occlusion, etc.

Our proposed Hand De-occlusion and Removal (HDR) framework is simple, flexible, and effective. Extensive experiments on the well-known InterHand2.6M benchmark [21] show that our method significantly outperforms the state-of-the-art 3D interacting hand pose estimation systems. Our framework builds upon the latest research progress of amodal segmentation [33], de-occlusion [38,36,19], and 3D single-hand pose estimation [39]. Note that, we do not perform complete comparisons with previous amodal segmentation, de-occlusion, and SHPE approaches. We also do not claim any algorithmic superiority concerning model architecture design. Because our aim is to propose a framework to solve the challenges of 3D interacting hand pose estimation. And designing powerful modules to improve the performance of amodal segmentation, de-occlusion, and SHPE is not the focus of this paper.

Our contributions are summarized as follows:

- We propose a novel Hand De-occlusion and Removal (HDR) framework to tackle the challenging task of 3D interacting hand pose estimation.
- We propose to explicitly handle the challenges of self-occlusion by hand de-occlusion and the homogeneous appearance ambiguity by distractor removal. To the best of our knowledge, we are the first to apply de-occlusion techniques to improve the downstream pose estimation accuracy.
- We propose the first large-scale synthetic Amodal InterHand Dataset (AIH) to settle the task of hand de-occlusion and removal. We envision that AIH will foster the development of the related research.

2 Related Work

2.1 Amodal Instance Segmentation and De-occlusion

Amodal Instance Segmentation. Unlike modal instance segmentation, which aims at assigning labels to visible parts of instances, amodal instance segmentation targets at producing the amodal (integrated) masks of each object instance

involving its occluded parts. Li and Malik [16] proposed the first amodal instance segmentation model which iteratively expands the bounding boxes and recomputes the heatmaps. Zhu *et al.* [40] proposed COCOA dataset for amodal instance segmentation and presented AmodalMask model as the baseline. Zhan *et al.* [36] propose a method to reason about the underlying occlusion ordering and recover the invisible parts in a self-supervised manner.

De-occlusion. De-occlusion aims at recovering the appearance content of the invisible occluded parts. SeGAN [6] adopts a residual network based model for mask completion and inferring the appearance of the invisible parts of indoor objects. Yan *et al.* [34] presented an iterative multi-task framework for amodal mask completion and de-occlusion of vehicles. Zhou *et al.* [38] built upon a well-known inpainting approach [19] and proposed to reason about the occluded regions and recover the appearance content of humans. Baek *et al.* [2] presents a weakly-supervised method to adapt from hand-object domain to single hand-only domain. However, its image generation module and the pose estimator are deeply coupled together, limiting its generalization ability to adapt to different hand pose estimators and resulting in low-quality restored image.

Our approach differs from previous works in three major aspects. First, previous works mostly focus on improving the quality of image content recovery, while we aim to improve the performance of the downstream task, *i.e.* 3D interactive hand pose estimation. Second, compared with common rigid objects, recovering the appearance content of the interacting hands is more challenging because of larger pose variations, severe hand-hand occlusion, and self-similar appearance of hands and fingers. Third, besides de-occlusion, our proposed HDR framework also performs distracting hand removal to reduce the ambiguities caused by the homogeneous appearance of hands.

2.2 Monocular RGB-based Hand Pose Estimation

Isolated hand pose estimation. RGB-based single (isolated) hand pose estimation has made significant progress in the past few years. Zimmermann *et al.* [41] introduced one of the first deep learning models to estimate hand poses from monocular RGB images. It first uses HandSegNet to localize hand regions, then uses PoseNet to estimate 2D hand poses, and finally maps 2D poses into 3D space. Iqbal *et al.* [11] proposed to encode hand joint locations with 2.5D heatmap representation to address the depth ambiguity problems and improve localization precision. Spurr *et al.* [29] proposed a VAE-based generative model to regress 3D hand joint locations. Zhou *et al.* [39] proposed to fully exploit non-image MoCap data to improve model generalization and robustness. Recently, many works also attempt to estimate 3D hand meshes from monocular RGB images. Most of them [1,4,35] are model-based, which train a convolutional neural network to estimate the MANO parameters [26]. Others are model-free, which directly regress 3D vertices of the human hand using mesh convolution [15], graph neural networks [5], or transformers [18].

Interacting hand pose estimation. Most works conduct interacting two-hand pose estimation by utilizing multi-view RGB images [3,9], depth data [23,22,30],

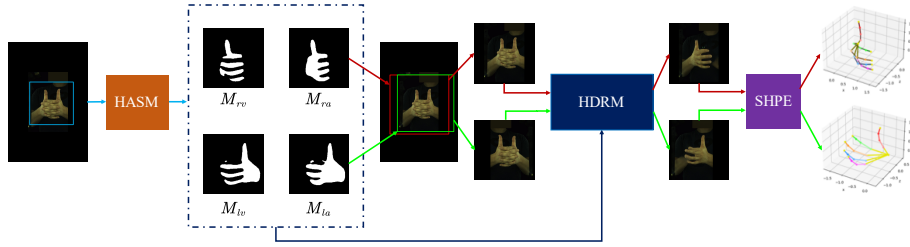


Fig. 2. Illustration of our Hand De-occlusion and Removal (HDR) framework for the task of 3D interacting hand pose estimation. We first employ **HASM** (Hand Amodal Segmentation Module) to segment the amodal and modal masks of the left and the right hand in the image. Given the predicted masks, we locate and crop the image patch centered at each hand. Then, for every cropped image, the **HDRM** (Hand De-occlusion and Removal Module) recovers the appearance content of the occluded part of one hand and removes the other distracting hand simultaneously. In this way, the interacting two-hand image is transformed into a single-hand image, and can be easily handled by **SHPE** (Single Hand Pose Estimation) to get the final 3D hand poses.

and tracking strategy [23,28,31]. Only a few existing works have considered estimating 3D poses of two hands from a single RGB image, which is challenging due to severe occlusion and close interactions. Lin *et al.* [17] employed a synthetic egocentric hand dataset to learn to estimate two-hand poses from a single RGB image. Moon *et al.* proposed a large-scale interacting hand dataset, termed InterHand2.6M dataset [21], and designed the InterNet model to predict 2.5D hand poses. Zhang *et al.* [37] designed a hand pose-aware attention module to address the self-similarity ambiguities and leveraged a context-aware cascaded refinement module to improve pose accuracy. Kim *et al.* [13] introduced an end-to-end trainable framework to jointly perform interacting hand pose estimation. Rong *et al.* [27] presented a two-stage framework to generate precise 3D hand poses and meshes with minimal collisions from monocular single RGB images. Fan *et al.* [7] proposed DIGIT (DIsembiGuating hands in InTeraction) to explicitly leverage the per-pixel probabilities to reduce the ambiguities caused by self-similarity of hands.

In this work, we empirically show that existing hand pose estimators often suffer from extreme self-occlusions and appearance ambiguity. To this end, we propose a novel Hand De-occlusion and Removal (HDR) framework to explicitly handle these two challenges, which significantly outperforms prior arts.

3 Method

3.1 Overview

As shown in Fig. 2, we propose a three-stage framework for interactive hand pose estimation. The first stage segments the complete and visible part for both



Fig. 3. Illustration of the HDRNet. The input of HDRNet includes 4 kinds of data: (a) the image erased on occluded portion of the right hand I_D , (b) the modal mask of the right hand M_{rv} , (c) the image erased on redundant portion of the distracting hand I_R , and (d) the modal mask of background M_{bv} . HDRNet recovers the appearance content of the occluded parts and inpaints the distracting hand to avoid ambiguity.

two hands. The second stage recovers the RGB values of the occluded hand and the background behind the distracting hand at the same time. The third stage predicts the 3D pose of each hand separately.

3.2 Hand Amodal Segmentation Module (HASM)

As shown in Fig. 2, given an interacting two-hand image, we first obtain the modal and amodal masks of both hands using the Hand Amodal Segmentation Module (HASM). We simply adapt the off-the-shelf instance segmentation model, *i.e.* SegFormer [33], to fit in our two-hand amodal segmentation tasks. Specifically, we increase the number of decode heads from one to four to predict four kinds of segmentation masks, namely the right hand amodal mask M_{ra} , the right hand visible mask M_{rv} , the left hand amodal mask M_{la} and the left hand visible mask M_{lv} . These segmentation masks contain (1) spatial localization information to roughly localize the left/right hand, and (2) rich cues about the occluded regions for de-occlusion and the distractor regions for removal.

We apply the binary cross entropy losses $\mathcal{L}_{BCE}(\cdot)$ to supervise the segmentation model. The final segmentation loss functions are formulated as follows:

$$\begin{aligned} \mathcal{L}_{HAS} = & \mathcal{L}_{BCE}(M_{ra}, M_{ra}^*) + \mathcal{L}_{BCE}(M_{lv}, M_{lv}^*) + \\ & \mathcal{L}_{BCE}(M_{la}, M_{la}^*) + \mathcal{L}_{BCE}(M_{rv}, M_{rv}^*), \end{aligned} \quad (1)$$

where M_{ra} , M_{lv} , M_{la} , and M_{rv} are predicted segmentation masks; M_{ra}^* , M_{lv}^* , M_{la}^* , and M_{rv}^* are the corresponding ground-truth masks.

3.3 Hand De-occlusion and Removal Module (HDRM)

Hand De-occlusion and Removal Module (HDRM) aims at transforming a previously challenging case of hand-hand interactions into a common single-hand

case, which can be easily solved by an off-the-shelf single-hand pose estimator. Specifically, given amodal and modal masks, *De-occlusion* is responsible for recovering the appearance content or RGB values of the occluded regions, while *Removal* targets at inpainting the distracting regions in the image, reducing the ambiguities caused by the homogeneous appearance of two hands.

For clarity, in the following sections, we will focus on the right hand only and regard the left hand as the distractor. Note that, the left-hand centered image can be flipped horizontally before performing hand de-occlusion, removal, and pose estimation, thus following the same pipeline.

As shown in Fig. 2, for the right hand, we first use the amodal mask M_{ra} to locate the right hand. Then we crop the original image and the segmentation masks at the center of the right hand. The newly cropped image and masks are denoted as I_s^{crop} , M_{ra}^{crop} , M_{rv}^{crop} , M_{la}^{crop} and M_{lv}^{crop} respectively. We will omit the superscript *crop* in subsequent sections for simplicity.

We use M_D to denote the region where the target hand is occluded by the other hand and M_R to denote the region where the distracting hand occupies. They are computed as follows:

$$\begin{aligned} M_D &= M_{ra} \cdot (1 - M_{rv}), \\ M_R &= (1 - M_{ra}) \cdot M_{lv}. \end{aligned} \quad (2)$$

I_D and I_R are the original image I_s erased by the mask M_D and M_R respectively. They can inform the HDRNet where to focus and how to inpaint these two regions with partial convolution [19]. In addition, the modal mask of the right hand M_{rv} and the modal mask of the background M_{bv} point out where the HDRNet can refer to for de-occlusion and removal respectively. Formally, I_D , I_R and M_{bv} are computed as follows:

$$\begin{aligned} I_D &= I_s \cdot (1 - M_D), \\ I_R &= I_s \cdot (1 - M_R), \\ M_{bv} &= (1 - M_{ra}) \cdot (1 - M_{la}). \end{aligned} \quad (3)$$

I_D , M_{rv} , I_R and M_{bv} are concatenated together as the input, as shown in Fig. 3. HDRNet then uses these data to recover the appearance content of the occluded parts and inpaints the distracting hand to avoid ambiguity. For model architecture choice, we follow [38,36] to adopt the network of Liu *et al.* [19] and further improve it by adding a few transformer blocks [32]. The transformer block enhances image feature interactions, enlarges the receptive fields, and focuses more on important image regions. Finally, the HDRNet outputs a recovered image I_o . We follow [38] to employ an image discriminator [12] D to enhance the image recovery quality through adversarial training. The loss function of HDRNet is as follows:

$$\begin{aligned} \mathcal{L}_{HDR} &= \lambda_1 (\mathbb{E}_{I_o} [\log(1 - D(I_o))] + \mathbb{E}_{I_o^*} [\log(D(I_o^*))]) + \\ &\quad \lambda_2 \mathcal{L}_{l1}(I_o, I_o^*) + \lambda_3 \mathcal{L}_{prec}(I_o, I_o^*) + \lambda_4 \mathcal{L}_{style}(I_o, I_o^*), \end{aligned} \quad (4)$$

where $\mathcal{L}_{prec}(\cdot)$ denotes the perceptual loss [8], and $\mathcal{L}_{style}(\cdot)$ denotes the style loss [19]. I_o is the recovered image, while I_o^* is its corresponding ground truth. λ_1 , λ_2 , λ_3 , and λ_4 are hyper-parameters to balance the losses.

3.4 3D Single Hand Pose Estimation (SHPE)

Our de-occlusion and removal framework can be applied to any off-the-shelf pose estimators. However, designing a more powerful hand pose estimation network architecture is not the focus of this paper. In this work, we choose the DetNet of MinimalHand [39] as our baseline SHPE for its simplicity and good performance. MinimalHand [39] comprises two modules, *i.e.* DetNet and IKNet. DetNet predicts the 2D and 3D hand joint positions. IKNet then takes as input the predicted 3D hand joint positions and maps them to the joint angles. In our implementation, we simply discard the IKNet and re-train the DetNet on the InterHand2.6M dataset [21]. The loss function of SHPE is as follows:

$$\mathcal{L}_{SHPE} = \mathcal{L}_{heat} + \mathcal{L}_{loc} + \mathcal{L}_{delta} + \mathcal{L}_{reg}, \quad (5)$$

where \mathcal{L}_{heat} is the 2D heatmap loss. \mathcal{L}_{loc} and \mathcal{L}_{delta} are location map loss and delta map loss respectively. \mathcal{L}_{reg} is a ℓ_2 weight regularizer to avoid overfitting. Please refer to Zhou *et al.* [39] for more training details.

4 Amodal InterHand (AIH) Dataset

Existing amodal perception datasets [40,24,10] mostly focus on amodal segmentation of common objects (*e.g.*, vehicles, buildings, and indoor objects). To the best of our knowledge, there exists no dataset that targets at amodal segmentation and appearance content recovery of interactive hands. To fill in this blank, we synthetically generate the first large-scale Amodal InterHand dataset, namely AIH dataset. We envision that the proposed dataset will boost the related research, *e.g.* amodal perception, de-occlusion, and hand pose estimation.

Our AIH dataset is constructed based on the well-known InterHand2.6M V1.0 dataset [21]. As shown in Fig. 4, our proposed Amodal InterHand (AIH) dataset consists of two parts: AIH_Syn and AIH_Render. In total, AIH dataset consists of about 3 million images, where AIH_Syn contains 2.2M samples and AIH_Render contains over 0.7M samples. AIH_Syn is generated by simple 2D image-level copy and paste, *i.e.* copy the left single-hand image and paste it on the right single-hand image; AIH_Render is obtained by rendering the textured interacting hand mesh to the image plane. Both AIH_Syn and AIH_Render contain the amodal and modal segmentation masks as well as the appearance content ground-truths.

AIH_Syn We first get the hand mesh with the ground-truth MANO parameters of the single-hand samples from the InterHand2.6M V1.0 dataset, and then project it into the 2D image plane to get the amodal segmentation mask. Then, we filter out some bad samples in which MANO parameters or the corresponding image are not valid. As a result, we get over 250K cropped single-hand images

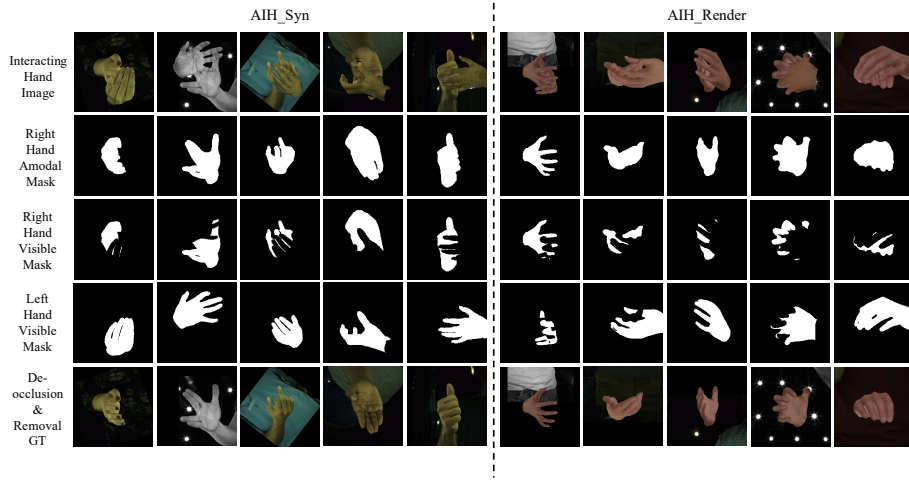


Fig. 4. Visualization of our proposed Amodal InterHand (AIH) dataset. AIH_Syn is obtained by simple 2D copy-and-paste, while AIH_Render is generated by rendering the textured 3D interacting hand mesh to the image plane.

with masks for each side hand. To generate the interacting two-hand samples, we randomly pick two hands with similar texture from both sides. Then we crop the left hand region given its amodal mask and paste it on the right hand image. Random scaling, rotation, and color jittering are applied to increase diversity.

AIH_Render Although AIH_Syn provides plenty of amodal data, such 2D level copy-and-paste can not generate mutual occlusion cases which are very common for interacting hands. Therefore, based on MANO parameters of InterHand2.6M V1.0 dataset, we decorate the corresponding hand mesh with random hand texture [25], augment it with random translation and rotation, and finally render it to a random background image from the dataset.

5 Experiments

5.1 Implementation Details

All experiments are conducted on 8 NVIDIA Tesla V100 GPUs. Training mini-batch size is set as 48 and Adam [14] is adopted for model parameter tuning. **Details of HASM.** Our HASM is trained for 200k iterations with a learning rate of 2.5×10^{-3} following the training setting of SegFormer [33]. **Details of HDRM.** Our HDRM has an input resolution of 256×256 . And the loss weights are set as $\lambda_1 = 0.1$, $\lambda_2 = 3.0$, $\lambda_3 = 0.1$, $\lambda_4 = 250.0$. We first train HDRM with ground-truth masks for 100k iteration, and then fine-tune it with segmentation masks for another 100k iteration. The learning rate of these two stages are 1.5×10^{-3} and 1×10^{-3} respectively. **Details of SHPE.** Our SHPE

has an input resolution of 256×256 . We train the network for 300k iterations with an initial learning rate of 1×10^{-3} . The learning rate is decayed to 1×10^{-4} and 1×10^{-5} at the 100k and 200k iterations respectively. Other training settings are kept the same as those of MinimalHand [39].

5.2 Datasets and Evaluation Metrics

Datasets. The experiments are conducted on InterHand2.6M V1.0 [21] dataset and Tzionas dataset [30]. *InterHand2.6M V1.0 dataset* [21] is a publicly available large-scale realistic pose estimation dataset for two-hand interactions. The dataset provides RGB images with semi-automatically annotated 3D poses, and MANO [26] parameters obtained from NeuralAnnot [20]. In the experiments, we follow the common practice [21] to use the downsized 512×334 image resolution at 5 frames-per-second (FPS) version of the released dataset. Following the official configurations, the dataset is split into three branches, namely ‘H’ for the human annotation branch, ‘M’ for the machine annotation branch, and ‘ALL’ for all data. The ‘M’ branch data contains many unseen poses and more diverse sequences, which makes it more similar to real-world scenarios. Moreover, we notice that the ‘H’ branch data contains missing or incomplete mesh annotations. In the experiments, we majorly conduct experiments on the ‘M’ branch, but also report the results on ‘ALL’ branch for comparisons. To focus on the interacting hands, the original dataset [21] divides the whole dataset (IH26M-ALL) into single-hands subset (IH26M-SH) and interacting-hands subset (IH26M-IH). Following [27], we further select samples from the original “IH26M-IH” test set, and generate a more challenging subset called “IH26M-Inter”. “IH26M-Inter” contains samples with more than 30 valid ‘ground-truth’ 3D hand keypoints. Since InterHand2.6M [21] is captured in a lab environment, its background diversity is relatively limited. To evaluate the generalization ability, we perform qualitative and quantitative experiments on the *Tzionas dataset* [30]. Since Tzionas dataset does not provide a separate training set, we directly use it as the testing set to evaluate the model trained on the InterHand2.6M dataset [21].

Evaluation Metrics. For InterHand2.6M [21], we report 3D Mean Per Joint Position Error (MPJPE). MPJPE is defined as the mean Euclidean distance between ground truth and predicted 3D joint locations, calculated after aligning the root joint for each left and right hand separately. The measurements are reported in millimeters (mm). For Tzionas dataset [30], we follow the common practice [21,4,13] to use 2D end point error (EPE) for evaluation.

5.3 Comparisons with state-of-the-art methods

We compare with previous state-of-the-art pose estimation methods on the ‘ALL’ branch and the ‘machine_annot (M)’ branch of InterHand2.6M V1.0 dataset [21]. MPJPE (mm) is adopted to evaluate the 3D hand pose estimation accuracy. For fair comparisons, the AIH dataset is only used to train HASM and HDRM for amodal segmentation and de-occlusion. No pose annotations in AIH are used to train the pose estimator (SHPE). Table 1 summarizes the experimental results.

Table 1. Comparisons with state-of-the-art methods on the ‘ALL’ branch and the ‘machine_annot (M)’ branch of InterHand2.6M V1.0 Dataset. MPJPE (mm) is adopted to evaluate the 3D joint estimation accuracy. The results marked with ‘*’ are from [27].

Methods	InterHand2.6M - ALL branch				InterHand2.6M - M branch		
	IH26M-SH	IH26M-IH	IH26M-ALL	IH26M-Inter	IH26M-SH	IH26M-IH	IH26M-ALL
*Boukhayma <i>et al.</i> [4]	-	-	27.14	31.46	-	-	-
*Pose2Mesh [5]	-	-	27.10	32.11	-	-	-
*BiHand [35]	-	-	25.10	28.23	-	-	-
*Rong <i>et al.</i> [27]	-	-	17.12	20.66	-	-	-
DIGIT [7]	-	14.27	-	-	-	-	-
InterNet [21]	12.16	16.02	14.21	18.04	12.52	18.04	15.28
HDR (Ours)	8.51	13.12	10.97	14.74	8.52	14.98	11.74

We first compare performances of three single-hand methods, *i.e.* Boukhayma *et al.* [4], Pose2Mesh [5] and BiHand [35]. Our approach significantly outperforms all the state-of-the-art single-hand approaches. On the “IH26M-ALL” split, compared with BiHand [35], our model reduces MPJPE from 25.10mm to 10.97mm, resulting in as much as 56% error reduction. And in the more challenging “IH26M-Inter” split, our approach obtains about 47% accuracy improvement. This shows existing single-hand pose estimators do not handle heavy hand-hand occlusions and are easily confused by the other distracting hand.

We also compare with recent two-hand pose estimation approaches, *i.e.* InterNet [21], Rong *et al.* [27], and DIGIT [7]. We show superior performance over these 3D interacting hand pose estimation systems. For example, our approach significantly improves upon Moon *et al.* [21]’s state-of-the-art results from 14.21mm to 10.97mm (about 23% error reduction) on the “IH26M-ALL” split. The clear performance gap validates the effectiveness of our framework. Overall, our approach consistently ranks the first across all evaluation protocols.

Table 2. Comparisons with state-of-the-art methods on Tzionas dataset [30]. The results of other algorithms are from [13]. 2D EPE is adopted to evaluate pose results.

Model	Boukhayma <i>et al.</i> [4]	Wang <i>et al.</i> [31]	InterNet [21]	Kim <i>et al.</i> [13]	SHPE	SHPE+HDR
EPE↓	12.91	13.31	17.61	12.42	14.88	8.70

We also follow [13] to report hand pose estimation results (EPE) on Tzionas dataset [30] in Table 2. Our method (SHPE+HDR in the table) significantly improves upon the baseline SHPE, and outperforms the prior arts.

5.4 Effect of Hand De-occlusion and Removal (HDR) Framework

As shown in Table 3, and Table 4, we conduct experiments on the ‘machine_annot (M)’ branch and the ‘ALL’ branch of InterHand2.6M V1.0 dataset respectively. We compare the results with or without using our HDR framework. We notice that the recent state-of-the-art single-hand pose estimation (SHPE) method (MinimalHand [39]) struggles with occlusions and appearance ambiguity in interacting hand scenarios (IH26M-IH). To tackle these challenges, we propose

Table 3. Effect of HDR Framework. Experiments are conducted on the ‘machine.annot (M)’ branch of InterHand2.6M V1.0 dataset. MPJPE (mm) is adopted to evaluate the 3D joint estimation accuracy.

Methods	Train (M, IH26M-SH)		Train (M, IH26M-SH +AIH)	
	IH26M-IH	IH26M-ALL	IH26M-IH	IH26M-ALL
SHPE [39]	40.98	25.78	32.27	21.66
+HDR (Ours)	25.45	17.98	24.59	17.80

Table 4. Effect of HDR Framework. Experiments are conducted on the ‘ALL’ branch of InterHand2.6M V1.0 dataset. MPJPE (mm) is adopted to evaluate the 3D joint estimation accuracy.

Methods	Train (ALL, IH26M-SH)		Train (ALL, IH26M-SH +AIH)	
	IH26M-IH	IH26M-ALL	IH26M-IH	IH26M-ALL
SHPE [39]	39.96	25.90	30.23	20.93
+HDR (Ours)	25.93	18.39	23.99	17.58

a novel Hand De-occlusion and Removal (HDR) framework to perform hand de-occlusion and distractor removal. In Table 3, we show that our approach significantly improves upon the SHPE baseline in interacting hand scenarios, *e.g.* from 40.98mm to 25.45mm (M, IH26M-IH). We find that adding AIH dataset for training will further improve the performance of SHPE, which validates the effect of AIH dataset. Experiments on the ‘ALL’ branch have a similar phenomenon.

5.5 Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of the key components of our approach on the ‘machine annot (M)’ branch of InterHand2.6M V1.0 dataset [21]. For fair comparisons, in all the ablation experiments, we use the same SHPE [39] trained on the IH26M-SH set.

Analysis of Hand De-occlusion and Removal (HDR) Module. There are two major challenges of hand pose estimation in interacting scenarios, *i.e.* severe self-occlusion, and ambiguity caused by the homogeneous appearance of hands. As shown in Table 5, #1, #2, #3, and #8, we conduct ablative experiments to quantitatively evaluate the effect of De-occlusion and Removal. Comparing #2 and #8, we observe that disabling ‘Removal’ will dramatically increase the MPJPE by 34.4%. Comparing #3 and #8, we see that disabling ‘De-occlusion’ increases the MPJPE by 9.5%. If we only apply SHPE [39] without HDR, the errors are further increased. These results clearly show that (1) state-of-the-art SHPE [39] is sensitive to self-occlusions and inter-hand ambiguities (2) HDRM is effective in handling the aforementioned two major challenges.

Analysis of Model Design Choices. We empirically validate the model design choice of HDRNet, especially *Discriminator* and the *Transformer* block. Discriminator is applied to enhance the image recovery quality by adversarial training. Comparing #4 and #8 in Table 5, we observe that although Discriminator helps in improving the quality of the recovered image, its influence on

Table 5. Ablation Studies. Experiments are conducted on the ‘machine_annot (M)’ branch of InterHand2.6M V1.0 dataset. We use MPJPE (mm) to evaluate the 3D joint estimation accuracy. Δ means the absolute (and relative) difference compared with our final model #8. ‘w/o’ is short for ‘without’.

	Methods	MPJPE (mm)	Δ
#1	SHPE [39] only	25.78	+7.80 (43.4%)
#2	w/o Removal	24.16	+6.18 (34.4%)
#3	w/o De-occlusion	19.69	+1.71 (9.5%)
#4	w/o Discriminator	18.11	+0.13 (0.7%)
#5	w/o Transformer Block	18.85	+0.87 (4.8%)
#6	AIH_Render only	18.10	+0.12 (0.7%)
#7	AIH_Syn only	18.35	+0.37 (2.1%)
#8	Ours	17.98	-

the final results is only marginal (0.7%). The Transformer block enhances image feature interactions, enlarges the receptive fields, and focuses more on important image regions. Comparing #5 and #8 in Table 5, we see that without using the Transformer block impacts the final results by a clear margin (4.8%).

Analysis of AIH_Syn and AIH_Render. Our proposed AIH dataset is composed of two subsets, namely AIH_Syn and AIH_Render. Both have their own advantages and disadvantages. AIH_Syn retains more detailed and realistic appearance features, while AIH_Render considers the inter-dependencies between two hands to avoid physically implausible configurations. Using a combination of these two sets to train the HDRNet will achieve the best performance. In Table 5, comparing #6, #7, and #8, we compare different training settings for HDRNet. We notice that it already achieves reasonably good results even if we only use one of the two sets. For example, using “AIH_Render only” to train HDRNet, we can achieve 18.10 MPJPE (mm), which is only marginally worse than the final model #8. We also empirically find that “AIH_Render” seems to have a larger impact on the final results than “AIH_Syn” does.

5.6 Time Complexity Analysis

We analyze the time cost on one Tesla P40 GPU in a single thread. On average, HASM, HDRM, and SHPE take 12.6 ms, 0.6 ms, and 34.0 ms per frame (including two hands) respectively. The time cost of HDRM (our major contribution) is only a small proportion of the total time cost (0.6 vs 47.2 ms).

5.7 Qualitative Results

In Fig. 5, we provide qualitative analysis on InterHand2.6M [21] and Tzionas dataset [30] to illustrate how HDR helps in handling severe hand-hand occlusion and the homogeneous appearance of hands. We see that HDR recovers the appearance in the occluded region and removes the distractor in challenging hand-hand occlusion cases. The results on Tzionas dataset [30] further validates the generalization ability of our proposed framework.

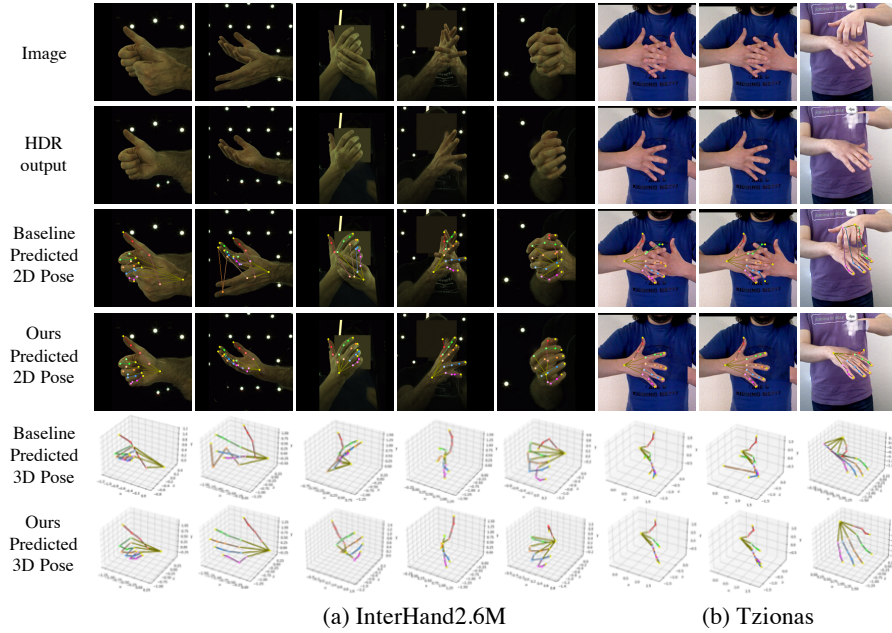


Fig. 5. Qualitive results of how HDR helps in handling severe hand-hand occlusion and appearance ambiguities of two hands. Best viewed in color.

6 Conclusions and Limitations

Interacting hand pose estimation is important but challenging due to severe hand-to-hand occlusion and ambiguity caused by the other distracting hand. In this paper, we propose to decompose the task into two relatively simple sub-tasks, *i.e.* (1) Hand De-occlusion and Removal (HDR), (2) Single Hand Pose Estimation (SHPE). Through HDR, we can simplify the case, which an off-the-shelf SHPE can handle. We empirically verified the effectiveness of our HDR framework on the InterHand2.6M and Tzionas dataset. Our limitations mainly lie in artifacts produced by HDRM. Improving the image recovery quality requires efforts in various research fields, which we will explore in the future.

Acknowledgement. We would like to thank Wentao Jiang, Wang Zeng, Neng Qian, Yumeng Hu, Lixin Yang, Yu Rong, Qiang Zhou and Jiayi Wang for their helpful discussions and feedback. Mengxiang Lin is supported by State Key Laboratory of Software Development Environment under Grant No SKLSDE 2022ZX-06. Ping Luo is supported by the General Research Fund of HK No.27208720, No.17212120, and No.17200622. Wanli Ouyang is supported by the Australian Research Council Grant DP200103223, Australian Medical Research Future Fund MRFAI000085, CRC-P Smart Material Recovery Facility (SMRF) – Curby Soft Plastics, and CRC-P ARIA - Bionic Visual-Spatial Prosthesis for the Blind.

References

1. Baek, S., Kim, K.I., Kim, T.K.: Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1067–1076 (2019) [4](#)
2. Baek, S., Kim, K.I., Kim, T.K.: Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020) [4](#)
3. Ballan, L., Taneja, A., Gall, J., Gool, L.V., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: Eur. Conf. Comput. Vis. pp. 640–653. Springer (2012) [4](#)
4. Boukhayma, A., Bem, R.d., Torr, P.H.: 3d hand shape and pose from images in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10843–10852 (2019) [4](#), [10](#), [11](#)
5. Choi, H., Moon, G., Lee, K.M.: Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In: Eur. Conf. Comput. Vis. (2020) [4](#), [11](#)
6. Ehsani, K., Mottaghi, R., Farhadi, A.: Segan: Segmenting and generating the invisible. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6144–6153 (2018) [4](#)
7. Fan, Z., Spurr, A., Kocabas, M., Tang, S., Black, M., Hilliges, O.: Learning to disambiguate strongly interacting hands via probabilistic per-pixel part segmentation. In: International Conference on 3D Vision (3DV) (2021) [5](#), [11](#)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2414–2423 (2016) [8](#)
9. Han, S., Liu, B., Cabezas, R., Twigg, C.D., Zhang, P., Petkau, J., Yu, T.H., Tai, C.J., Akbay, M., Wang, Z., et al.: Megatrack: monochrome egocentric articulated hand-tracking for virtual reality. ACM Trans. Graph. **39**(4), 87–1 (2020) [4](#)
10. Hu, Y.T., Chen, H.S., Hui, K., Huang, J.B., Schwing, A.G.: Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3105–3115 (2019) [8](#)
11. Iqbal, U., Molchanov, P., Gall, T.B.J., Kautz, J.: Hand pose estimation via latent 2.5 d heatmap regression. In: Eur. Conf. Comput. Vis. pp. 118–134 (2018) [4](#)
12. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1125–1134 (2017) [7](#)
13. Kim, D.U., Kim, K.I., Baek, S.: End-to-end detection and pose estimation of two interacting hands. In: Int. Conf. Comput. Vis. pp. 11189–11198 (2021) [5](#), [10](#), [11](#)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [9](#)
15. Kulon, D., Guler, R.A., Kokkinos, I., Bronstein, M.M., Zafeiriou, S.: Weakly-supervised mesh-convolutional hand reconstruction in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4990–5000 (2020) [4](#)
16. Li, K., Malik, J.: Amodal instance segmentation. In: Eur. Conf. Comput. Vis. pp. 677–693. Springer (2016) [4](#)
17. Lin, F., Wilhelm, C., Martinez, T.: Two-hand global 3d pose estimation using monocular rgb. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2373–2381 (2021) [5](#)
18. Lin, K., Wang, L., Liu, Z.: End-to-end human pose and mesh reconstruction with transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1954–1963 (2021) [4](#)

19. Liu, G., Reda, F.A., Shih, K.J., Wang, T.C., Tao, A., Catanzaro, B.: Image inpainting for irregular holes using partial convolutions. In: *Eur. Conf. Comput. Vis.* (2018) [3](#), [4](#), [7](#), [8](#)
20. Moon, G., Lee, K.M.: Neuralannot: Neural annotator for in-the-wild expressive 3d human pose and mesh training sets. *arXiv preprint arXiv:2011.11232* (2020) [10](#)
21. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: *Eur. Conf. Comput. Vis.* pp. 548–564. Springer (2020) [3](#), [5](#), [8](#), [10](#), [11](#), [12](#), [13](#), [18](#)
22. Mueller, F., Davis, M., Bernard, F., Sotnychenko, O., Verschoor, M., Otaduy, M.A., Casas, D., Theobalt, C.: Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM Trans. Graph.* **38**(4), 1–13 (2019) [4](#)
23. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the articulated motion of two strongly interacting hands. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 1862–1869. IEEE (2012) [4](#), [5](#)
24. Qi, L., Jiang, L., Liu, S., Shen, X., Jia, J.: Amodal instance segmentation with kins dataset. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3014–3023 (2019) [8](#)
25. Qian, N., Wang, J., Mueller, F., Bernard, F., Golyanik, V., Theobalt, C.: HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In: *Eur. Conf. Comput. Vis.* Springer (2020) [9](#)
26. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022) [4](#), [10](#)
27. Rong, Y., Wang, J., Liu, Z., Loy, C.C.: Monocular 3d reconstruction of interacting hands via collision-aware factorized refinements. In: *International Conference on 3D Vision* (2021) [5](#), [10](#), [11](#)
28. Smith, B., Wu, C., Wen, H., Peluse, P., Sheikh, Y., Hodgins, J.K., Shiratori, T.: Constraining dense hand surface tracking with elasticity. *ACM Trans. Graph.* **39**(6), 1–14 (2020) [5](#)
29. Spurr, A., Song, J., Park, S., Hilliges, O.: Cross-modal deep variational hand pose estimation. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 89–98 (2018) [4](#)
30. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. *Int. J. Comput. Vis.* **118**(2), 172–193 (2016) [4](#), [10](#), [11](#), [13](#), [18](#)
31. Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M.A., Casas, D., Theobalt, C.: Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM Trans. Graph.* **39**(6), 1–16 (2020) [5](#), [11](#)
32. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pvt2: Improved baselines with pyramid vision transformer. *Computational Visual Media* **8**(3), 1–10 (2022) [7](#)
33. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. *Adv. Neural Inform. Process. Syst.* **34** (2021) [3](#), [6](#), [9](#)
34. Yan, X., Wang, F., Liu, W., Yu, Y., He, S., Pan, J.: Visualizing the invisible: Occluded vehicle segmentation and recovery. In: *Int. Conf. Comput. Vis.* pp. 7618–7627 (2019) [4](#)
35. Yang, L., Li, J., Xu, W., Diao, Y., Lu, C.: Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In: *Brit. Mach. Vis. Conf.* (2020) [4](#), [11](#)
36. Zhan, X., Pan, X., Dai, B., Liu, Z., Lin, D., Loy, C.C.: Self-supervised scene de-occlusion. In: *IEEE Conf. Comput. Vis. Pattern Recog.* pp. 3784–3792 (2020) [3](#), [4](#), [7](#)

- 37. Zhang, B., Wang, Y., Deng, X., Zhang, Y., Tan, P., Ma, C., Wang, H.: Interacting two-hand 3d pose and shape reconstruction from single color image. In: Int. Conf. Comput. Vis. pp. 11354–11363 (2021) [5](#)
- 38. Zhou, Q., Wang, S., Wang, Y., Huang, Z., Wang, X.: Human de-occlusion: Invisible perception and recovery for humans. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3691–3701 (2021) [3](#), [4](#), [7](#)
- 39. Zhou, Y., Habermann, M., Xu, W., Habibie, I., Theobalt, C., Xu, F.: Monocular real-time hand shape and motion capture using multi-modal data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5346–5355 (2020) [3](#), [4](#), [8](#), [10](#), [11](#), [12](#), [13](#), [18](#)
- 40. Zhu, Y., Tian, Y., Metaxas, D., Dollár, P.: Semantic amodal segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1464–1472 (2017) [4](#), [8](#)
- 41. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: Int. Conf. Comput. Vis. pp. 4903–4911 (2017) [4](#)

Appendix

A Video Demo

To justify the generalization ability and the potential of our proposed method in real-world applications, we run our approach on several video clips from the Tzionas dataset [30]. Note that our models are only trained on InterHand2.6M V1.0 dataset [21] and our proposed AIH dataset. Tzionas dataset [30] is *unseen* during training.

In the video demo¹, we compare our approach with ‘Baseline’ which is a Single-Hand Pose Estimator (SHPE). For fair comparisons, both ‘Ours’ and ‘Baseline’ employ the same SHPE [39] trained on the ‘ALL’ branch of InterHand2.6M V1.0 dataset [21]. Note that this model is the same as the one used in Sec. 5.6 (Fig. 5) of the main paper. The pose results are directly obtained from the output of the SHPE model without temporal smoothing.

We first visualize the predicted amodal/visible mask of both hands. Given the segmentation mask, we obtain the corresponding single-hand box (‘red’ for the right hand, and ‘green’ for the left hand). We also demonstrate the results of Hand De-occlusion and Removal Module (HDRM). In order to tackle the severe hand-hand occlusion problem, HDRM applies the hand de-occlusion technique to recover the appearance (texture) in the occluded region. In the meanwhile, it also removes (inpaints) the distracting hand to handle the ambiguity caused by the homogeneous appearance of hands. Finally, our approach obtains better 3d hand pose estimation results.

Although the quality of the image recovery is satisfactory in most cases, there are still some problems in some difficult situations. For example, the boundary of the hand segmentation can be over-smoothed, leading to undesirable artifacts around the hand. This problem can be mitigated by applying an advanced amodal/visible mask segmentation model, which we will explore in the future.

B More Examples of AIH Dataset

In this section, we present more examples of our proposed Amodal InterHand (AIH) dataset. Our AIH dataset consists of two parts: AIH_Syn and AIH_Render. AIH_Syn is constructed by copy-and-paste while AIH_Render is constructed by rendering the textured interacting hand mesh to the image plane. As shown in Fig A1, both AIH_Syn and AIH_Render have great diversity in hand poses, textures, occlusion and interaction types.

¹ Our video demo can be downloaded from https://connecthkuhk-my.sharepoint.com/:v:/g/personal/js20_connect_hku_hk/EW_S3kZu97xPlMk_HQLAJVMBtizU48sGh4jXwvUuyugFRw?e=u2GB6I.



Fig. A1. Top: More examples of our proposed AIH_Syn dataset. **Bottom:** More examples of our proposed AIH_Render dataset.