Bringing Inputs to Shared Domains for 3D Interacting Hands Recovery in the Wild

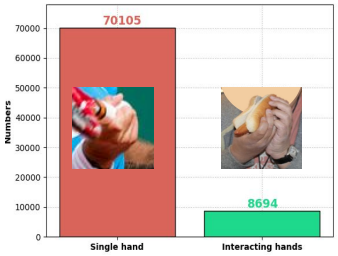
Gyeongsik Moon
Meta Reality Labs
mks0601@gmail.com

Abstract

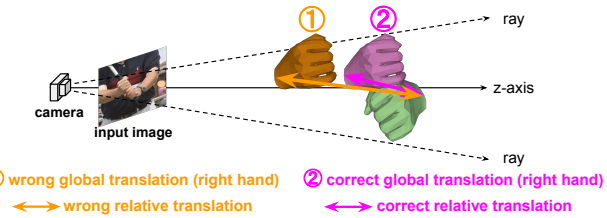
Despite recent achievements, existing 3D interacting hands recovery methods have shown results mainly on motion capture (MoCap) environments, not on in-the-wild (ITW) ones. This is because collecting 3D interacting hands data in the wild is extremely challenging, even for the 2D data. We present InterWild, which brings MoCap and ITW samples to shared domains for robust 3D interacting hands recovery in the wild with a limited amount of ITW 2D/3D interacting hands data. 3D interacting hands recovery consists of two sub-problems: 1) 3D recovery of each hand and 2) 3D relative translation recovery between two hands. For the first sub-problem, we bring MoCap and ITW samples to a shared 2D scale space. Although ITW datasets provide a limited amount of 2D/3D interacting hands, they contain large-scale 2D single hand data. Motivated by this, we use a single hand image as an input for the first sub-problem regardless of whether two hands are interacting. Hence, interacting hands of MoCap datasets are brought to the 2D scale space of single hands of ITW datasets. For the second sub-problem, we bring MoCap and ITW samples to a shared appearance-invariant space. Unlike the first sub-problem, 2D labels of ITW datasets are not helpful for the second sub-problem due to the 3D translation's ambiguity. Hence, instead of relying on ITW samples, we amplify the generalizability of MoCap samples by taking only a geometric feature without an image as an input for the second sub-problem. As the geometric feature is invariant to appearances, MoCap and ITW samples do not suffer from a huge appearance gap between the two datasets. The code is publicly available[1].

1. Introduction

3D interacting hands recovery aims to reconstruct a single person's interacting right and left hands in the 3D space. The recent introduction of a large-scale motion capture (MoCap) dataset [20] motivated many 3D interacting hands recovery methods [4, 6, 14, 25, 32].



(a) The number of single hand vs. interacting hands in MSCOCO



(b) The failure case of the 2D-based weak supervision for the 3D relative translation between two hands

Figure 1. (a) Only a very small amount of 2D interacting hands are available in ITW datasets despite a relaxed threshold of intersection-over-union (IoU). We consider two hands are interacting if the IoU between the two hands' boxes is bigger than 0.1. (b) The 2D-based weak supervision from ITW datasets often results in a wrong 3D relative translation (**the orange arrow**).

Although they have shown robust results on MoCap datasets, none of them explicitly tackled robustness on in-the-wild (ITW) datasets. Simply training networks on MoCap datasets and testing them on ITW datasets results in unstable results due to a huge domain gap between MoCap and ITW datasets. The most representative domain gap is an appearance gap. For example, images in InterHand2.6M [20] (IH2.6M) have black backgrounds and artificial illuminations, far from those of ITW datasets. The fundamental solution for this is collecting large-scale ITW data with 3D groundtruths (GTs); however, this is extremely challenging. For example, capturing 3D data requires tens of calibrated

[1] https://github.com/facebookresearch/InterWild

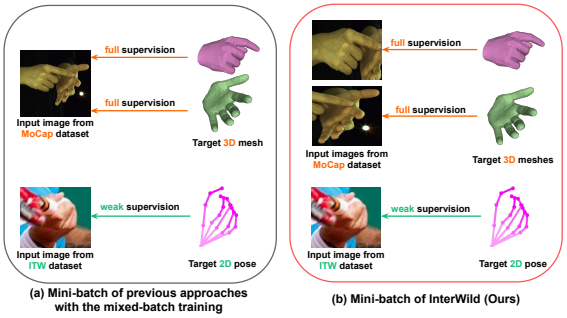


Figure 2. Mini-batch comparison for the first sub-problem (*i.e.*, estimation of separate 3D meshes of left and right hands).

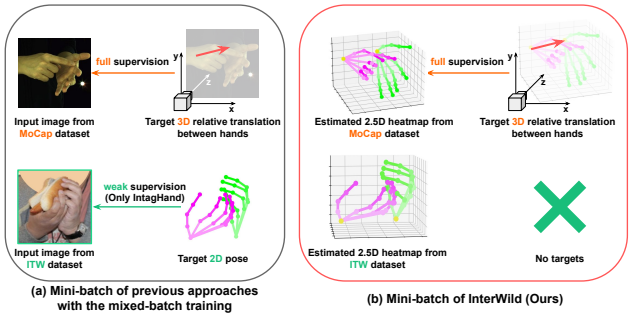


Figure 3. Mini-batch comparison for the second sub-problem (*i.e.*, estimation of 3D relative translation between two hands).

and synchronized cameras. Preparing such a setup at diverse places in the wild requires a huge amount of manual effort. Furthermore, collecting even large-scale ITW 2D interacting hand data with manual annotation is greatly challenging due to the severe occlusions and self-similarities. Due to such challenges, there is no large-scale ITW 2D/3D interacting hand dataset.

Nevertheless, ITW datasets provide large-scale 2D *single-hand* data, as shown in Fig. 1 (a). Utilizing such large-scale 2D single-hand data of ITW datasets can be an orthogonal research direction to the 2D/3D interacting hands data collection in the wild. Mixed-batch training is the most dominant approach to utilize 2D data of ITW datasets for the 3D human recovery [3, 11, 12, 17, 19, 24]. During the mixed-batch training, half of the samples in a mini-batch are taken from MoCap datasets and the rest of the samples from ITW datasets. The MoCap samples are fully supervised with 3D GTs, and the ITW samples are weakly supervised with 2D GTs. The ITW samples make networks exposed to diverse appearances, which leads to successful generalization to unseen ITW images. The 2D-based weak supervision is enabled by the MANO [23] hand model, which produces a 3D hand mesh from pose and shape parameters in a differentiable way. To be specific, 3D joint coordinates, extracted from the 3D mesh, are projected to the 2D space using an estimated 3D global translation (*i.e.*, 3D translation from the camera to the hand root joint) and fixed virtual camera intrinsics. Then, the projected 2D joint coordinates are supervised with the 2D GTs. In this way, the 2D GTs weakly supervise MANO parameters, which can make all vertices of the 3D mesh fit to the 2D GTs.

However, naively re-training networks of previous 3D interacting hands recovery methods [4, 6, 14, 25, 32] with the mixed-batch training does not result in robust results. 3D interacting hands mesh recovery consists of two sub-problems: 1) estimation of separate 3D right and left hands and 2) estimation of 3D relative translation between two hands. For the first sub-problem, previous works [4, 6, 14, 25, 32] take an image of two hands when hands are interacting, and an image of a single hand when hands are not interacting (Fig. 2 (a)). As ITW datasets mostly contain single-hand data (Fig. 1 (a)), most samples from ITW datasets contain a single hand during the mixed-batch training. The problem is that the images of two hands from MoCap datasets have very different 2D hand scale distribution compared to that of single-hand images from ITW datasets, as shown in Fig. 5. For example, when two hands are included in the input image, the 2D scale of each hand is much smaller than that from a cropped image of a single hand.

Unlike the first sub-problem, the second sub-problem hardly gets benefits from the 2D-based weak supervision from ITW datasets. Fig. 1 (b) shows the failure case of the 2D-based weak supervision. When the 3D global translation, estimated for the 2D-based weak supervision, is wrong (① in the figure), the 3D relative translation is supervised to be wrong one (**the orange arrow** in the figure). The wrong 3D global translation also can happen to the first sub-problem; however, the critical difference is that the 3D scale of the 3D relative translation (*i.e.*, output of the second sub-problem) is very weakly constrained, while the 3D scale of hands (*i.e.*, output of the first sub-problem) are strongly constrained by the shape parameter of MANO [23]. For example, we can place two hands close or far freely based on how they are interacting with each other, while the size of adults' hands is usually around 15 cm. As there is no such strong constraint to the relative translation, the relative translation can be an arbitrary value and is prone to wrong supervision (**the orange arrow** in the figure). Please note that estimating a 3D global translation from a single image involves a high ambiguity and can be often wrong as the camera position is not provided in the input image. In this regard, we observed that the 2D-based weak supervision for the 3D relative translation, used in IntagHand [14] (Fig. 3 (a)) deteriorates results, which is shown in the experimental section. However, without the 2D-based weak supervision of ITW datasets, the network is trained only on images of MoCap datasets like previous works [4,6,25,32] (Fig. 3 (a)), which results in generalization failure due to the appearance gap between MoCap and ITW datasets.

We present InterWild, a framework for 3D interacting hands mesh recovery in the wild. For the first sub-problem, InterWild takes a cropped single-hand image regardless of whether two hands are interacting or not, as shown in Fig. 2 (b). In this way, the 2D scales of interacting hands are normalized to those of a single hand. Such normalization brings all single and interacting hands to the shared 2D scale space; hence, large-scale single-hand data of ITW datasets can be much more helpful compared to the counterpart without the normalization.

For the second sub-problem, InterWild takes geometric features without images, as shown in Fig. 3 (b). In particular, the output of the second sub-problem (*i.e.*, 3D relative translation) is fully supervised only for MoCap samples to prevent the 2D-based weak supervision of ITW samples from deteriorating the 3D relative translation. The geometric features are invariant to appearances, such as colors and illuminations, which can reduce the huge appearance gap between MoCap and ITW datasets and bring samples from two datasets to a shared appearance-invariant space. Therefore, although the estimated 3D relative translation is supervised only on MoCap datasets and is not supervised on ITW datasets, our InterWild produces robust 3D relative translations on ITW datasets.

We show that our InterWild produces highly robust 3D interacting hand meshes from ITW images. As 3D interacting hands recovery in the wild is barely studied, we hope that ours can give useful insight into future works. For the continual study, we released our codes and trained models.

Our contributions can be summarized as follows.

- We present InterWild, a framework for the 3D interacting hands recovery in the wild.
- For the separate left and right 3D hands, InterWild takes a cropped single-hand image regardless of whether hands are interacting or not so that all hands are brought to a shared 2D scale space.
- For the 3D relative translation between two hands, InterWild takes only geometric features, which are invariant to appearances.

2. Related works

3D interacting hands recovery. Most of early works [1, 21, 22, 28, 29, 31] fit 3D hand models to geometric evidence, such as RGBD sequence [22], hand segmentation map [21], and dense matching map [31]. Recently, Moon *et al*. [18, 20] presented IH2.6M dataset, the first large-scale real-captured dataset that contains accurate GT 3D poses and meshes of interacting hands and a regression-based baseline model, InterNet. Motivated by IH2.6M and InterNet, several regression-based methods have been proposed, which perform better than the above fitting-based methods. Rong *et al*. [25] proposed a two-stage framework to minimize collisions between two hands. Zhang *et al*. [32] proposed a cascaded 3D interacting hand mesh estimation network, which sequentially refines 3D interacting hands. Kwon *et al*. [13] presented a baseline for recovering 3D meshes of two hands while interacting with objects. However, their system mostly focuses on the interaction between hands and objects, not between two hands. Li *et al*. [14] proposed a graph convolutional network for accurate 3D interacting hand reconstruction. Hampali *et al*. [6] proposed a Transformer [30]-based system that separates localization and identification of hand keypoints. Di *et al*. [4] presented a lightweight system for 3D interacting hand mesh recovery. In addition, there are several works [5, 10, 16] that recover only 3D hand joint locations without 3D meshes.

The above 3D interacting hand reconstruction methods [5, 10, 14, 16, 20, 25, 32] fail to produce robust results on ITW datasets, while ours can. There are two big differences. First, they take a two-hand image as an input when two hands are interacting (Fig. 2 (a)). On the other hand, ours take a single-hand image regardless of whether two hands are interacting (Fig. 2 (b)); hence, inputs are brought to the shared 2D scale space. Second, they estimate 3D relative translation between two hands using an image with the 2D-based weak supervision (Fig. 3 (a)). On the other hand, ours estimates the relative translation only from geometric features (Fig. 3 (b)), which are invariant to appearances, without the 2D-based weak supervision.

Reducing appearance gap with geometric features. Several 3D body and hand mesh estimation methods have used geometric features to reduce the appearance gap between MoCap and ITW datasets. Pose2Mesh [2] and Song *et al*. [26] take 2D body joint coordinates as an input to predict a 3D human body mesh. Zhou *et al*. [34] predict a 3D single-hand mesh from 3D single-hand joint coordinates. Zhang *et al*. [33] utilizes body part UVI map for the 3D body mesh recovery. Our InterWild is the first work that estimates robust 3D relative translation between two hands from geometric features.

3. InterWild

Fig. 4 shows the overall pipeline of our InterWild, which consists of DetectNet, SHNet, and TransNet. DetectNet detects hands from the input image. Then, SHNet, a network for a single hand, takes each detected hand image as an input and outputs 3D mesh and 2.5D pose of each hand. The 2.5D poses of the right and left hands are passed to TransNet, which outputs 3D relative translation between two hands. The final 3D interacting hands are obtained by adding the 3D relative translation to the 3D mesh of the left hand. DetectNet and SHNet follow architectures of Pose2Pose [17]. Please refer to the supplementary material

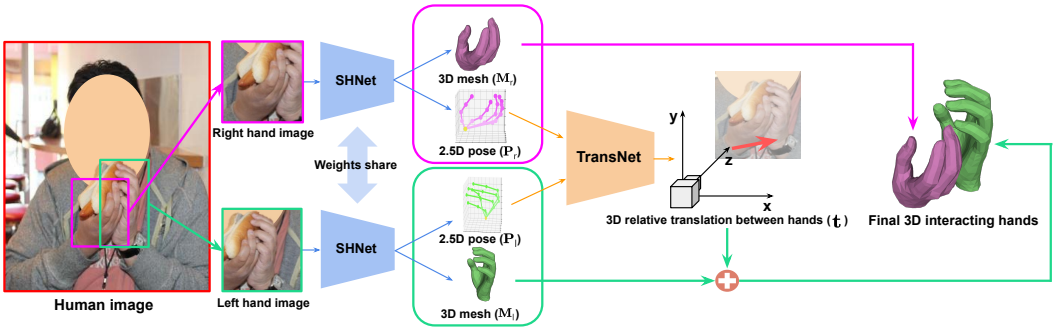


Figure 4. The overall pipeline of the proposed InterWild. From the hand boxes, obtained by DetectNet, we crop and resize the hand area from the high-resolution human image. Each right and left hand image is fed to SHNet, which produces 3D mesh and 2.5D heatmap. Next, TransNet takes the 2.5D heatmap of two hands to produce the 3D relative translation between two hands. Final 3D interacting hand meshes are obtained by adding the 3D relative translation to the left hand mesh. For simplicity, we do not visualize DetectNet.

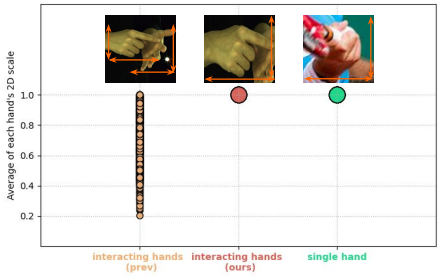


Figure 5. Average of each hand's width and height, where the width and height are normalized with the size of the input image. We extended all hands' boxes to set their aspect ratio to 1 before calculating the scales. Yellow: Each hand in the two-hand image of IH2.6M [20] when hands are interacting (Previous approach). Brown: Each hand in the single-hand image of IH2.6M [20] when hands are interacting (Ours). Green: Each hand in the single-hand image of MSCOCO [15].

for their detailed architectures.

3.1. SHNet

Input: an image of a single hand. SHNet takes a single-hand image regardless of whether two hands are interacting or not, while previous methods take images with two hands when hands are interacting, as shown in Fig. 2. Hence, 2D scales of interacting hands are normalized to those of a single hand. Fig. 5 shows that when we crop images to contain two hands when hands are interacting (*i.e.*, previous methods. Yellow in the figure), 2D scales of each hand in two-hand images have a very different distribution compared to those of each hand in single-hand images (Green in the figure). On the other hand, when we crop images to contain a single hand regardless of whether hands are interacting (*i.e.*, ours. Brown in the figure), 2D scales of each hand in two-hand images have almost the same distribution compared to those of each hand in single-hand images (Green in the figure). Such an analysis justifies our design of SHNet to take a single cropped hand.

The single-hand image is cropped and resized from the high-resolution human image using predicted boxes from the DetectNet. Before cropping the hands, we double the width and height of boxes to prevent hands from missing and provide more surrounding context to SHNet. The left hand image is horizontally flipped to the right hand; therefore, the input image always represents a right hand image. The right hand and flipped left hand images are concatenated in the batch dimension and processed in a parallel way by the SHNet. By taking the right hand and flipped left hand images, SHNet can focus only on learning to process right hand images, which can relieve the burdens of learning to process both right and left hand images. Also, such flipping is helpful when the two hands are severely interacting so that boxes of two hands are largely overlapped. For example, let us imagine that most of the left hand is occluded by the right hand. Then, images from the left and right hand boxes would contain almost the same right hand. By flipping the left hand image, the right hand in the original left hand image changes to the left hand. We train SHNet to ignore the left hand in the input image and produce a 3D hand mesh of only the right hand in the input image. Therefore, the output from the flipped left hand image is a 3D hand mesh of the occluded right hand, which is originally the occluded left hand. The effectiveness of this flipping is shown in the experimental section.

Output: 3D mesh and 2.5D pose of each hand. Using the network architecture of Pose2Pose [17], our SHNet outputs 3D mesh and 2.5D pose [27] of each hand. We flip back the outputs of the flipped left hand image. We denote the 3D mesh of left and right hands by $\mathbf{M}_{\text{l}}$ and $\mathbf{M}_{\text{r}}$, respectively. Each 3D mesh is obtained by forwarding the estimated pose and shape parameters to a MANO [23] layer. We subtract 3D meshes from their 3D root joint locations so that the 3D meshes are in the root joint-relative space. In addition, we denote the 2.5D pose of left and right hands by $\mathbf{P}_{\text{l}} \in \mathbb{R}^{J\times 3}$ and $\mathbf{P}_{\text{r}} \in \mathbb{R}^{J\times 3}$, respectively. J indicates the number of single-hand joints. The 2.5D pose encodes hand joint locations in 2.5D space. The x- and y-axis of the jth 2.5D pose

represent pixel coordinates of the jth joint, where the pixel space is defined in the input image of SHNet (*i.e.*, single-hand image). The z-axis is defined in the root joint-relative depth space.

3.2. TransNet

Fig. 6 shows the overall pipeline of TransNet, a network to predict 3D relative translation between two hands.

Input: 2.5D poses of two hands. TransNet takes 2.5D poses of two hands, while previous methods take images with two hands, as shown in Fig. 3. The 2.5D poses of two hands are from SHNet, which are denoted by $\mathbf{P}_{\text{r}}$ and $\mathbf{P}_{\text{l}}$. Before forwarding them, we apply 2D affine transformations to $\mathbf{P}_{\text{r}}$ and $\mathbf{P}_{\text{l}}$, which transform the input space of SHNet (*i.e.*, an image of a single hand) to a union of two-hand boxes space (*i.e.*, an image of two hands). By warping them to the union hand box space, we can get a relative 2D scale and translation between two hands in the 2D pixel space. Based on such relative 2D information and pose information, TransNet predicts the 3D relative translation.

For example, when xy distance of two hands' 2.5D pose is small, (x, y) of the 3D relative translation are close to zero. Also, when one hand takes smaller area in input xy space, that hand might have larger depth; however, not always true as hands are deformable. When a hand is in neutral pose and the other one is in fist pose, their 3D relative depth can be zero although the hand with fist pose takes smaller area. Hence, pose is necessary to determine z of the 3D relative translation. Please note that the 2D affine transformations do not affect the depths of each 2.5D pose; hence, the depths of each 2.5D pose still represent the root joint-relative depths of each hand. We denote the transformed $\mathbf{P}_{\text{r}}$ and $\mathbf{P}_{\text{l}}$ by $\mathbf{P}'_{\text{r}}$ and $\mathbf{P}'_{\text{l}}$, respectively.

The 2.5D pose of the right hand $\mathbf{P}'_{\text{r}}$ and left hand $\mathbf{P}'_{\text{l}}$ are converted to 2.5D Gaussian heatmaps by making a Gaussian blob around the coordinates. By converting coordinates to heatmaps, we can exploit the strong feature extraction power of ResNet [8] as ResNet takes tensor inputs, not vector inputs. Then, we concatenate the 2.5D Gaussian heatmap of two hands in a channel dimension, denoted by $\mathbf{H} \in \mathbb{R}^{2J \times D \times H \times W}$. D, H, and W represent the depth, height, and width of the 2.5D heatmap, respectively, and we set them to 64.

Output: 3D relative translation between two hands. We predict the 3D relative translation between two hands $\mathbf{t} \in \mathbb{R}^3$ from the 2.5D Gaussian heatmap $\mathbf{H}$. We pass $\mathbf{H}$ to ResNet-18 [8], which produces a feature map $\mathbf{F} \in \mathbb{R}^{C \times H/8 \times W/8}$. $C = 512$ represents the channel dimension of $\mathbf{F}$. We use the original ResNet-18 after dropping the first convolutional block to reduce the downsampling and the last fully-connected layers. As the 3D relative translation represents a 3D relative location of the left wrist from the right wrist, extracting useful wrist information is a key

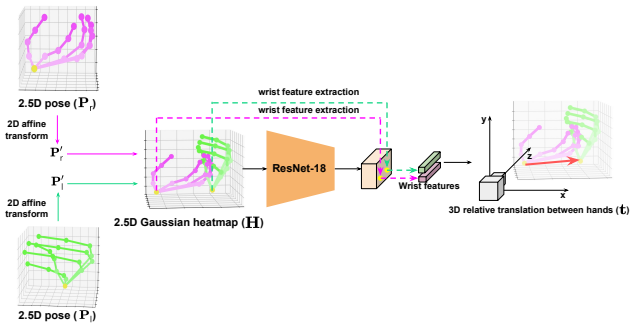


Figure 6. The overall pipeline of TransNet. It applies a 2D affine transformation to the 2.5D pose of each left and right hand to bring them to the union hand box space of the original input image. Then, wrist features are extracted for the 3D relative translation estimation.

for accurate 3D relative translation. However, most existing methods [5, 20, 25, 32] perform global average pooling (GAP) to the last feature map of backbones and pass the output to several fully connected layers. As GAP simply averages the spatial dimension, it might not be effective to capture useful wrist information. Instead, we perform a bilinear interpolation at the 2D positions of left and right wrists in $\mathbf{F}$, where the 2D wrist positions are from $\mathbf{P}'_{\text{r}}$ and $\mathbf{P}'_{\text{l}}$. The extracted wrist features are concatenated with the 2D wrist coordinates and fed to a linear layer, which finally produces 3D relative translation $\mathbf{t}$. The effectiveness of our wrist feature extraction compared to previous GAP-based approaches is shown in the experimental section.

3.3. Final outputs and loss functions

The final 3D interacting hand meshes consist of 1) the 3D mesh of the right hand $\mathbf{M}_{\text{r}}$ and 2) a summation of the 3D mesh of the left hand $\mathbf{M}_{\text{l}}$ and the 3D relative translation $\mathbf{t}$. We train InterWild in an end-to-end manner by minimizing $L1$ distance between predicted and GT boxes, MANO parameters, 3D joint coordinates, and 3D relative translation. Please note that the 3D relative translation is only supervised by MoCap datasets as ITW datasets do not provide GTs.

4. Experiments

4.1. Datasets

Train sets. IH2.6M [20] and the whole-body version of MSOCO [9, 15] are used for the training. During the training, the mini-batch consists of half-IH2.6M and half-MSCOCO samples.

Test sets. As our primary goal is 3D interacting hand mesh recovery *in the wild*, we use Hands In Action dataset (HIC) [29] as our main test set. HIC [29] contains single and interacting hand sequences captured with an RGBD camera. HIC provides 3D GT meshes of hands [7], fitted to

Inputs of SHNet	HIC [29]	IH2.6M [20]
Two-hand image	29.80 / 35.86	11.36 / 13.20
Single-hand image (Ours)	**15.65 / 15.70**	**11.12 / 13.01**

Table 1. MPVPE comparisons between SHNets that take an image of 1) two hands and 2) a single hand as an input when hands are interacting. Both settings take a single hand image when hands are not interacting. The left and right numbers for each setting represent errors from single and interacting hand sequences, respectively.

Settings	HIC [29]	IH2.6M [20]
Without flip / shared weights	15.80 / 17.36	12.54 / 18.21
Without flip / separated weights	**15.34** / 16.96	11.48 / 13.77
With flip / shared weights (Ours)	15.65 / **15.70**	**11.12 / 13.01**

Table 2. MPVPE comparisons of SHNets that take 1) left and right hands and 2) flipped left hand and right hand. 'Shared weights' indicates that SHNet's weights are shared for left and right hand inputs, while 'separated weights' indicates that the weights are not shared. The left and right numbers for each setting represent errors from single and interacting hand sequences, respectively.

a 3D point cloud. Although HIC is captured in an indoor environment, it contains images with much more diverse and realistic appearances compared to those of IH2.6M. Also, as we do not use HIC during the training, its appearances are not exposed to the network; hence, we believe the test performance of networks on HIC represents generalizability to unseen appearances, necessary for the 3D interacting hand mesh recovery in the wild. Additionally, we report errors on IH2.6M as it is one of the representative datasets for the 3D interacting hand mesh recovery although it is a MoCap dataset. Qualitative results are shown on MSCOCO, which is the most widely used ITW dataset due to its diverse appearances.

4.2. Evaluation metrics

MPJPE and MPVPE. Mean per-joint position error (MPJPE) and mean per-vertex position error (MPVPE) evaluate 3D joint and mesh vertex positions, respectively. It represents the average 3D joint and mesh vertex distance (mm) between the predicted and GT, after aligning those with a root joint translation. MPJPE and MPVPE are used to measure 3D errors of 3D mesh of each hand.

MRRPE. Mean relative-root position error (MRRPE) evaluates 3D relative translation between two hands. It calculates a 3D distance (mm) between the predicted and GT right hand root-relative left hand root position.

4.3. Ablation study

SHNet: Effectiveness of taking an image of a single hand. Table 1 shows that when hands are interacting, taking a single hand image (Fig. 2 (b)) produces lower SHNet's errors compared to taking a two-hands image (Fig. 2 (a)), especially on HIC. This shows that our approach to taking a single hand image when hands are interacting is especially helpful in the wild. The reason for our superior result is that when SHNet takes a single hand image, large-scale 3D interacting hand data of MoCap dataset [20] and large-scale 2D single hand data of in-the-wild dataset [15] are brought to a shared 2D scale space. On the other hand, when taking images of two hands when hands are interacting like previous works [5, 10, 14, 16, 20, 25, 32], the 2D scale of each hand has a very different distribution, as shown in Fig. 5; hence, learning to process inputs with a very different distribution can be a burden to SHNet. The small gap of IH2.6M is because such a burden can be relieved by large-scale interacting hand datasets, such as IH2.6M. We observed that reducing the number of IH2.6M's interacting hand samples makes the same tendency of HIC, which supports our claim. However, collecting large-scale interacting hand data in the wild is greatly challenging; even collecting 2D data requires a huge amount of effort due to the severe self-similarity and occlusions. We effectively relieve such data collection burden by bringing the two datasets to a shared 2D scale space. Both settings take a single hand when hands are not interacting, and the first setting takes a two-hand image when hands are interacting. For the setting that takes an image of two hands, we doubled the output part of SHNet to estimate both left and right hands at the same time.

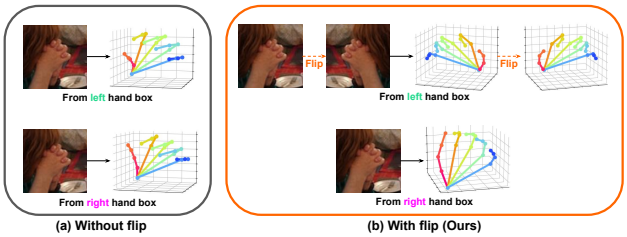


Figure 7. Comparison between 2.5D hand pose from SHNets that take (a) an image of left and right hand images and (b) an image of *flipped left hand* and right hand. The result from the flipped left hand is flipped back.

SHNet: Effectiveness of flipping left hand images. Table 2 shows that flipping the left hand is necessary for SHNet's accurate results when hands are interacting. Without the flipping, SHNet can take almost the same single hand image when hands are severely overlapped, as shown in Fig. 7 (a). As input images are almost the same, SHNet outputs almost the same two 3D hands for both left and right hand images. On the other hand, when we pass a flipped left hand image, SHNet can differentiate the two images. As SHNet is trained to ignore all left hands and recover only right hands, this can be seen as an *implicit de-occlusion* of left hands from the input image. Fig. 7 (b) shows that even when two hands are severely overlapped, our SHNet can recover correct 3D hands.

Instead of flipping, one can double the output part of SHNet and train SHNet to output left and right hands at

Inputs of TransNet	weak sup.	HIC [29]	IH2.6M [20]
Img.	✗	206.83	27.67
	✓	215.35	35.72
Img. + 2.5D hm.	✗	54.36	**27.19**
	✓	58.53	33.15
2D hm.	✗	38.64	31.51
	✓	51.19	35.51
2.5D hm.	✗**(Ours)**	**31.35**	29.29
	✓	61.05	33.91

Table 3. MRRPE comparisons of TransNets that take various inputs and are trained without and with the 2D-based weak supervision. The hm. represents a heatmap.

Settings	HIC [29]	IH2.6M [20]
GAP	39.85	**29.14**
All joint features	48.99	31.57
Wrist features (Ours)	**31.35**	29.29

Table 4. MRRPE comparisons between TransNet that outputs the 3D relative translation with various feature extraction settings.

the same time. In this way, when the input hand images are almost the same due to the severe overlap, the output part of SHNet is trained to distinguish left and right hands. However, we observed that our flipping results in better results than this variant. We think this is because, in our setting, the handedness of the input image is normalized to the right hand, while the variant is not. Such normalization can relieve the burden of SHNet. Please note that a combination of 'with flip' and 'separated weights' is impossible as flipping normalizes the handedness.

TransNet: Effectiveness of the geometric inputs. Table 3 shows that taking 2.5D heatmaps as an input of TransNet (Fig. 3 (b)) achieves the lowest MRRPE on HIC and comparable results on IH2.6M. Our 2.5D heatmap achieves better results than the 2D heatmap due to the additional depth information of each hand. An interesting result is that using an image as an input (the first and second rows) achieves good results on IH2.6M, but bad results on HIC. Such a setting with the image input is similar to previous methods [5,14,20,25,32], while IntagHand [14] additionally uses segmentation and DensePose.

There are two reasons for this setting's high MRRPE on HIC. First, when the 2D-based weak supervision is disabled, the huge appearance gap between ITW and IH2.6M is the main reason. Without the weak supervision, TransNet is supervised only on IH2.6M. Images of MoCap datasets, including IH2.6M, have monotonous colors with artificial illuminations, which are far from those of ITW images. On the other hand, we use pure geometric features (*i.e.*, 2.5D heatmap) as it is invariant to appearances. Thanks to the invariance, our TransNet successfully generalizes to ITW datasets although it is trained only on IH2.6M. Second, when the 2D-based weak supervision is enabled, the weak supervision deteriorates the relative translation due to the scale ambiguity of the relative translation. A detailed analysis of the 2D-based weak supervision is provided below.

Methods	HIC [29]		IH2.6M [20]	
	MPVPE	MRRPE	MPVPE	MRRPE
IHMR [25]	30.76 / 46.38	119.64	15.35 / 18.53	33.39
Zhang *et al.* [32]	23.53 / 31.79	110.25	11.76 / 14.17	31.56
IntagHand [14]	18.83 / 27.31	52.46	11.18 / 13.49	29.31
InterWild (Ours)	**15.65 / 15.70**	**31.35**	**11.12 / 13.01**	**29.29**

Table 5. Comparison of our InterWild and 3D interacting hand mesh estimation methods.

Methods	MPJPE	MRRPE
AIH [16]	76.83 / 36.05	N/A
InterWild (Ours)	**16.00 / 16.17**	**31.35**

Table 6. Comparison of our InterWild and 3D interacting hand pose estimation methods on HIC.

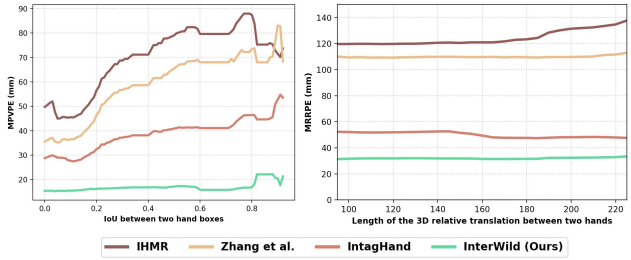


Figure 8. Comparison with previous methods [14, 25, 32] on HIC. For each x-axis value, denoted by τ, MPVPE and MRRPE of y-axis are calculated from samples whose x-axis values are larger than τ.

TransNet: Bad effect of the 2D-based weak supervision. Table 3 shows that introducing the 2D-based weak supervision for the estimation of the 3D relative translation between two hands, similar to IntagHand [14], deteriorates MRRPE for all inputs of TransNet and for both evaluation benchmarks. This is because, unlike the 3D scales of hands that are strongly constrained with the shape parameter of MANO, 3D scales of the 3D relative translation are very weakly constrained. For example, we can put two hands near or far based on how they are interacting with each other, while the size of adults' hands is usually around 15 cm. Without such a strong constraint, the 3D relative translation can be an arbitrary value due to the wrong 3D global translation. Fig. 1 (b) shows that when the 3D global translation is wrong (① in the figure), the 3D relative translation is supervised to be wrong one (**the orange arrow** in the figure). Please refer to the supplementary material for how we introduced the 2D-based weak supervision.

TransNet: Effectiveness of the wrist features. Table 4 shows that our wrist feature-based estimation of 3D relative translation achieves better results than the previous widely used GAP [5, 20, 25, 32]. This is because GAP averages the entire spatial domain; therefore, its output is not guaranteed to contain essential wrist information, necessary for the 3D relative translation between two hands. On the other hand, we explicitly extract wrist features, which produces better results. Interestingly, using features of all joints performs worst. We think that this is because features from all joints

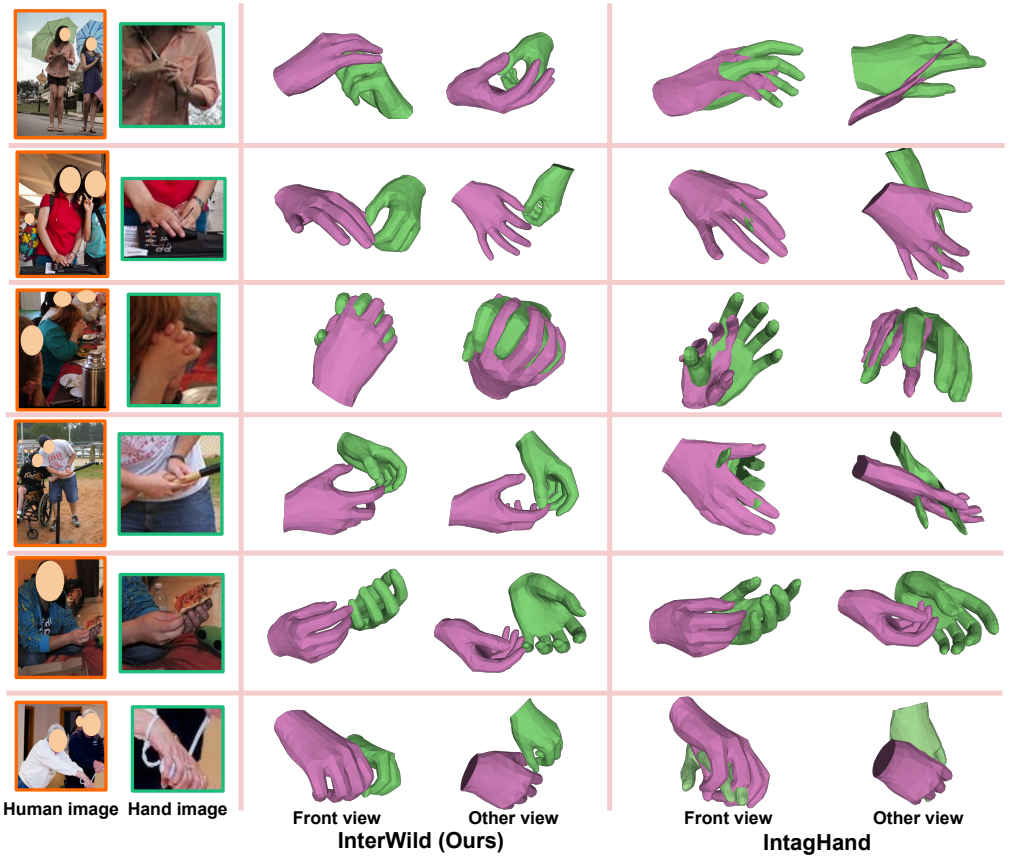


Figure 9. Qualitative comparison between our InterWild and IntagHand [14] on MSCOCO validation set. Ours detects hand boxes from the human images, while IntagHand takes the hand images using GT boxes.

contain too much unnecessary information, which gives a burden to the regressor.

4.4. Comparison with state-of-the-art methods

Table 5 and Fig. 8 show that ours outperforms all existing 3D interacting hand mesh recovery methods on HIC and IH2.6M datasets. It is noteworthy that the MPVPE and MRRPE gap is especially large on HIC, which shows the robustness of InterWild to ITW environments. This is because of our domain-sharing approach, depicted in Fig. 2 and 3. Table 6 additionally demonstrates the superiority of our InterWild. AIH [16] used additional synthetic datasets for the robust results on unseen appearances. However, such synthetic datasets are still built on top of the IH2.6M dataset, which has a severe appearance gap from that of ITW datasets. On the other hand, ours effectively reduces the domain gap by bringing inputs of SHNet and TransNet to shared domains, which results in a strong performance on ITW datasets. Finally, Fig. 9 visually demonstrates that ours successfully recovers 3D meshes from ITW images, while previous state-of-the-art method [14] fails to.

Publicly released models of previous works in Table 5 are trained only on IH2.6M, and their training codes are not available. Therefore, we reproduced their networks based on their testing codes and re-trained all of their networks on IH2.6M and MSCOCO like ours for a fair comparison. We will verify our reproduce results in the supplementary material. For the evaluation, we do not align the scale with GTs following Moon *et al*. [20]. We found that Keypoint Transformer [6] produces bad results when MSCOCO is incorporated in the training set as it requires 3D GTs, which does not exist in MSCOCO, for the camera parameter loss function. All previous methods use GT hand boxes during the inference following their settings, while ours uses predicted hand boxes from DetectNet.

5. Conclusion

We present InterWild, a framework for the 3D interacting hand mesh recovery in the wild. InterWild effectively reduces the domain gap between MoCap and ITW datasets. To this end, it takes an image of a single hand regardless of whether hands are interacting for the estimation of 3D meshes of left and right hands. In addition, it takes geometric features for the estimation of 3D relative translation between two hands. Integrating our InterWild with whole-body 3D human mesh estimation methods can be a promising future research direction.

References

[1] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 3

[2] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3D human pose and mesh recovery from a 2D human pose. In *ECCV*, 2020. 3

[3] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *CVPR*, 2022. 2

[4] Xinhan Di and Pengqian Yu. LWA-HAND: Lightweight attention hand for interacting hand reconstruction. In *ECCVW*, 2022. 1, 2, 3

[5] 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 *3DV*, 2021. 3, 5, 6, 7

[6] 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 *CVPR*, 2022. 1, 2, 3, 8

[7] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 5

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[9] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *ECCV*, 2020. 5

[10] Dong Uk Kim, Kwang In Kim, and Seungryul Baek. End-to-end detection and pose estimation of two interacting hands. In *ICCV*, 2021. 3, 6

[11] Muhammed Kocabas, Chun-Hao P Huang, Otmar Hilliges, and Michael J Black. PARE: Part attention regressor for 3D human body estimation. In *ICCV*, 2021. 2

[12] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 2

[13] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 3

[14] 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 *CVPR*, 2022. 1, 2, 3, 6, 7, 8

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 4, 5, 6

[16] 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 *ECCV*, 2022. 3, 6, 7, 8

[17] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Accurate 3D hand pose estimation for whole-body 3D human mesh estimation. In *CVPRW*, 2022. 2, 3, 4

[18] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3D human mesh training sets. In *CVPRW*, 2022. 3

[19] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-Lixel prediction network for accurate 3D human pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 2

[20] 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 *ECCV*, 2020. 1, 3, 4, 5, 6, 7, 8

[21] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG*, 2019. 3

[22] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Tracking the articulated motion of two strongly interacting hands. In *CVPR*, 2012. 3

[23] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied Hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017. 2, 4

[24] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshop*, 2021. 2

[25] Yu Rong, Jingbo Wang, Ziwei Liu, and Chen Change Loy. Monocular 3D reconstruction of interacting hands via collision-aware factorized refinements. In *3DV*, 2021. 1, 2, 3, 5, 6, 7

[26] Jie Song, Xu Chen, and Otmar Hilliges. Human body model fitting by learned gradient descent. In *ECCV*, 2020. 3

[27] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 4

[28] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM TOG*, 2016. 3

[29] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 3, 5, 6, 7

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3

[31] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: real-time tracking of 3D hand interactions from monocular RGB video. *ACM TOG*, 2020. 3

[32] 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 *ICCV*, 2021. 1, 2, 3, 5, 6, 7

[33] Hongwen Zhang, Jie Cao, Guo Lu, Wanli Ouyang, and Zhenan Sun. Learning 3D human shape and pose from dense body parts. *TPAMI*, 2020. 3

[34] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 3