2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) | 978-1-6654-8739-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/CVPRW56347.2022.00257

Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation

Gyeongsik Moon[1] Hongsuk Choi[1] Kyoung Mu Lee[1,2]

[1]Dept. of ECE & ASRI, [2]IPAI, Seoul National University, Korea

{mks0601,redarknight,kyoungmu}@snu.ac.kr

Abstract

Whole-body 3D human mesh estimation aims to reconstruct the 3D human body, hands, and face simultaneously. Although several methods have been proposed, accurate prediction of 3D hands, which consist of 3D wrist and fingers, still remains challenging due to two reasons. First, the human kinematic chain has not been carefully considered when predicting the 3D wrists. Second, previous works utilize body features for the 3D fingers, where the body feature barely contains finger information. To resolve the limitations, we present Hand4Whole, which has two strong points over previous works. First, we design Pose2Pose, a module that utilizes joint features for 3D joint rotations. Using Pose2Pose, Hand4Whole utilizes hand MCP joint features to predict 3D wrists as MCP joints largely contribute to 3D wrist rotations in the human kinematic chain. Second, Hand4Whole discards the body feature when predicting 3D finger rotations. Our Hand4Whole is trained in an end-to-end manner and produces much better 3D hand results than previous whole-body 3D human mesh estimation methods. The codes are available here [1].

1. Introduction

Whole-body 3D human mesh estimation aims to localize mesh vertices of all human parts, including body, hands, and face, simultaneously in the 3D space. By combining 3D mesh of all human parts, we can understand not only human body pose and shape but also human intention and feeling through hand poses and facial expressions. The challenges of this problem are small image sizes and complicated articulations of hands, and smoothly connecting estimated 3D body and hands meshes.

Existing whole-body 3D human mesh estimation systems [4,5,25,34] consist of body, hands, and face networks. As hands and face take small areas in the human image, the previous works crop and resize the hands and face images to higher resolutions. Then, they process the human, hands, and face images using body, hand, and face networks, respectively. Their body, hand, and face networks perform global average pooling (GAP) to the extracted image features and predict 3D body, hands, and face, respectively, which consist of 3D joint rotations and other parameters (*e.g.*, body shape and facial expressions). The 3D body, hands, and face are passed to SMPL-X layer [24] for the final whole-body 3D mesh.

3D hands recovery consists of the 3D wrist and finger rotation predictions. Accurate 3D hands recovery is a key

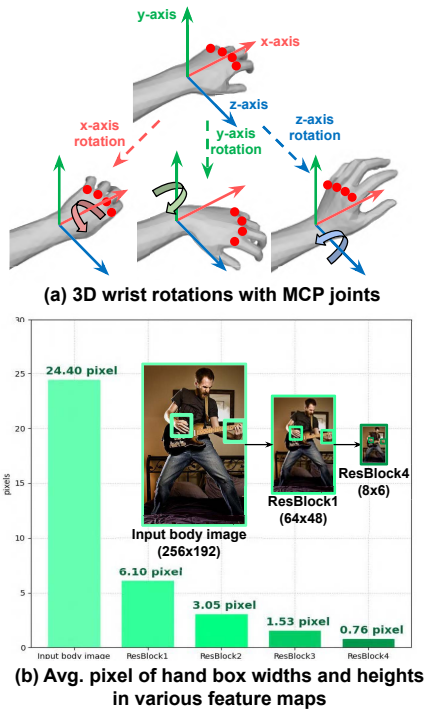

Figure 1. (a) Four hand MCP joints (red circles) provide essential 3D wrist rotation information as they are child nodes of a wrist in hand kinematic chain. (b) The average of width and height (pixel) of hand boxes in MSCOCO [12]. The detailed finger information is almost missing at the output of the standard backbone, ResNet (less than 1 pixel).

[1]https://github.com/mks0601/Hand4Whole_RELEASE

978-1-6654-8739-9/22/$31.00 ©2022 IEEE
DOI 10.1109/CVPRW56347.2022.00257

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

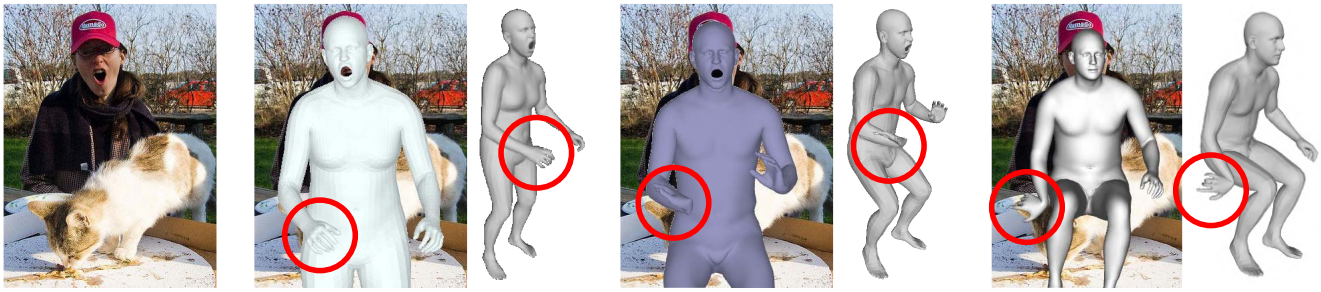

Figure 2. Qualitative results comparison between the proposed Hand4Whole, ExPose [4], and FrankMocap [25] when hands are invisible. Taking a body feature is necessary for plausible 3D wrist rotations.

for the whole-body 3D human mesh estimation; however, previous works [4, 5, 25, 34] produce inaccurate 3D hands due to two reasons. First, for the 3D wrist rotation, they have not carefully considered the human kinematic chain. In the human kinematic chain, the wrist connects the human body and fingers' root joints (*i.e.*, hand MCP joints); therefore, utilizing human body and MCP joint information is necessary for accurate 3D wrist rotation. In particular, the MCP joint information provides the necessary cue to determine 3D wrist rotations, as shown in Figure 1 (a). Furthermore, the body information provides overall body posture, which can make the predicted 3D wrist rotation be anatomically plausible and smoothly connected with the body, even when hands are occluded or truncated, as shown in Figure 2. However, none of the previous works used a combination of the body feature and MCP joint features for the 3D wrist rotations. All of them used one of the body feature [34], hand feature [4, 25], and their combination [5], where the hand feature contains lots of unnecessary information, such as finger information, while the MCP joint feature contains essential 3D wrist rotation information. In particular, the finger information can hurt the 3D wrist rotation accuracy as the fingers often move independently to the wrists and have highly complicated articulations.

Second, for the 3D finger rotations, previous works [5, 34] used body features in addition to hand features. The body feature mainly contains much unnecessary information, such as body and backgrounds, while having very coarse hand information due to the small sizes of hands, as shown in Figure 1 (b). Therefore, such unnecessary information can corrupt the 3D finger rotation accuracy.

To resolve the above limitations, we present Hand4Whole, a whole-body 3D human mesh estimation system that produces much better 3D hands in the whole-body 3D mesh. Hand4Whole has two stronger points than previous works. First, we design Pose2Pose, a 3D positional pose-guided 3D rotational pose prediction framework. Here, the 3D positional and rotational pose represent 3D joint positions and rotations, respectively. In contrast to previous works [4, 5, 25] that vectorize the image feature by performing GAP, ours extract joint features by our positional pose-guided pooling (PPP). The joint features enable Pose2Pose to understand different semantic information of each joint, while the previously used vectorized image feature only provides only an instance-level understanding. Using Pose2Pose, Hand4Whole utilizes a combination of the body feature and eight MCP joint features of two hands for the 3D wrist rotations. The body feature makes the 3D wrist rotations anatomically plausible when hands are occluded or truncated, as shown in Figure 2. Also, from an anatomical point of view, the eight hand MCP joint features provide essential 3D wrist rotations among thirty finger joints of two hands, as shown in Figure 1 (a). By combining both features, Hand4Whole produces much more accurate 3D wrist rotations.

Second, we discard body features when predicting 3D finger rotations. As a result, we can prevent the coarse hand information from corrupting the 3D finger rotation accuracy. Although we do not use the body feature for 3D finger rotation, the 3D hands can be anatomically plausible and smoothly connected with body joints as our 3D wrist rotations already consider the body network's feature. In particular, the wrists are root nodes of a human hand kinematic chain, which determine global 3D rotations of fingers. Our Hand4Whole is trained in an end-to-end manner and significantly outperforms previous whole-body 3D human mesh estimation methods.

Our contributions can be summarized as follows.

- Our Hand4Whole produces much more accurate 3D hands, which consists of the 3D wrist and finger rotations. Using our Pose2Pose, Hand4Whole utilizes both body and hand MCP joint features for accurate 3D wrist rotation and smooth connection between 3D body and hands.
- In addition, we discard body features when predicting 3D finger rotations.
- Hand4Whole is trained in an end-to-end manner and largely outperforms previous whole-body 3D human

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

mesh estimation methods.

2. Related works

Although there have been remarkable progress in 3D body-only [3, 14, 16, 18, 20, 21, 28], hand-only [2, 7, 17, 35], and face-only [26] mesh estimation, there have been few attempts to simultaneously recover the 3D body, hands, and face. Several early works are based on an optimization-based approach, which fits a 3D human model to the 2D/3D evidence. Joo *et al*. [13] fits their human models (*i.e.*, Frank and Adam) to 3D human joints coordinates and point clouds in a multi-view studio environment. Xiang *et al*. [30] extended Joo *et al*. [13] to the single RGB case. SMPLif-X [24] and Xu *et al*. [32] fit their human models, SMPL-X and GHUM, respectively, to 2D human joint coordinates. As the above optimization-based methods can be slow and prone to noisy evidence, a regression-based approach is presented recently. Recently, several neural network-based methods have been proposed. All of them consist of body, hand, and face networks, where each takes a human image, hand-cropped image, and face-cropped image, respectively. The three networks predict SMPL-X parameters for a final whole-body 3D human mesh. ExPose [4] is the earliest regression-based approach, which utilizes the three separated networks. FrankMocap [25] consists of body and hand networks, and their hand prediction is attached to 3D body results. Zhou *et al*. [34] utilize 3D joint coordinates for the 3D joint rotation prediction. PIXIE [5] introduced a moderator to predict 3D joint rotations.

Compared to the recent whole-body 3D human mesh estimation methods, our Hand4Whole has three clear novel contributions. First, we design Pose2Pose as the main module of our Hand4Whole. Pose2Pose enables Hand4Whole to understand different semantic information of each joint by introducing joint features, while previous works [4,5,25] provides only an instance-level feature to their system by performing the GAP. Second, Hand4Whole uses both body and hand MCP joint features for the 3D wrist rotations. On the other hand, ExPose [4] and FrankMocap [25] predict 3D wrist rotation only from a hand feature without a body feature. The absence of the body feature when predicting the 3D wrist rotations results in implausible 3D wrist rotations when hands are occluded or truncated, as shown in Figure 2. Zhou *et al*. [34] predict 3D wrist rotations only from a body feature without a hand feature. They suffer from inaccurate 3D wrist rotations as they do not utilize MCP joint features. PIXIE [5] combines body and hand features by a moderator to predict 3D wrist rotations. However, their hand feature contains lots of unnecessary information, such as finger information, which can hurt the 3D wrist rotation accuracy as fingers often move independently to the wrists. Third, Hand4Whole discards the body feature when predicting 3D finger rotations. In contrast, Zhou *et al*. [34] and PIXIE [5] uses the body feature, which contains very coarse hand information due to the small pixel size of the hands.

3. Hand4Whole

Figure 3 shows the overall pipeline of the proposed Hand4Whole for whole-body 3D human mesh estimation. It consists of BodyNet, HandNet, and FaceNet, which take cropped and resized human, hands, and face images, respectively. The outputs of each network are fed to SMPL-X [24] layer to obtain the final whole-body 3D human mesh.

3.1. Pose2Pose

To start, we describe Pose2Pose, a 3D positional pose (*i.e.*, 3D joint positions)-guided 3D rotational pose (*i.e.*, 3D joint rotations) prediction framework. Figure 4 shows the overall pipeline of Pose2Pose when it is used for the 3D body. Pose2Pose enables the proposed Hand4Whole to understand different semantic information of each joint. We use Pose2Pose as the main module of our BodyNet and HandNet. Pose2Pose consists of two stages, described in the following.

3D joint coordinates estimation. Pose2Pose firsts predicts the 3D joint coordinates $\mathbf{P} \in \mathbb{R}^{J\times 3}$ from a human image. J denotes the number of joints. x- and y-axis of $\mathbf{P}$ are in pixel space, and z-axis of it is in root joint (*i.e.*, pelvis for the body and wrist for the hand)-relative depth space. To this end, ResNet-50 [10] extracts 2048-dimensional image feature map $\mathbf{F}$ from the input image. We use ResNet after removing the GAP and fully-connected layer of the last part of the original ResNet. Then, a 1-by-1 convolutional layer predicts 3D heatmaps of human joints $\mathbf{H}$ from $\mathbf{F}$, where $\mathbf{H}$ has the same height and width as $\mathbf{F}$. To make 3D heatmaps from the 2D feature map, the 1-by-1 convolutional layer changes the channel dimension from 2048 to $8J$. Then, we reshape the $8J$-dimensional feature map to J dimensional 3D heatmap whose depth size is 8. The 3D joint coordinates $\mathbf{P}$ is obtained from $\mathbf{H}$ by the soft-argmax operation [27] in a differentiable way.

Positional pose-guided pooling (PPP). Positional pose-guided pooling extracts joint features from the 2D feature map $\mathbf{F}$. To this end, we apply a 1-by-1 convolution to $\mathbf{F}$ to change its channel dimension from 2048 to 512. Then, we perform a bilinear interpolation to the output of the convolution at (x, y) position of 3D joint coordinates $\mathbf{P}$, which produce joint features $\mathbf{F}^P \in \mathbb{R}^{J\times 512}$. Please note that as (x, y) coordinates of $\mathbf{P}$ are in pixel space, we use them directly for the interpolation without a 3D-to-2D projection. The joint features provide different contextual information of each joint, useful for the 3D joint rotation [8].

3D joint rotations estimation. For the 3D joint rotation θ, we concatenate the joint features $\mathbf{F}^P$ and 3D joint coordinates $\mathbf{P}$ and flatten it, denoted by $\mathbf{v} \in \mathbb{R}^{512J+3J}$. Then, $\mathbf{v}$

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

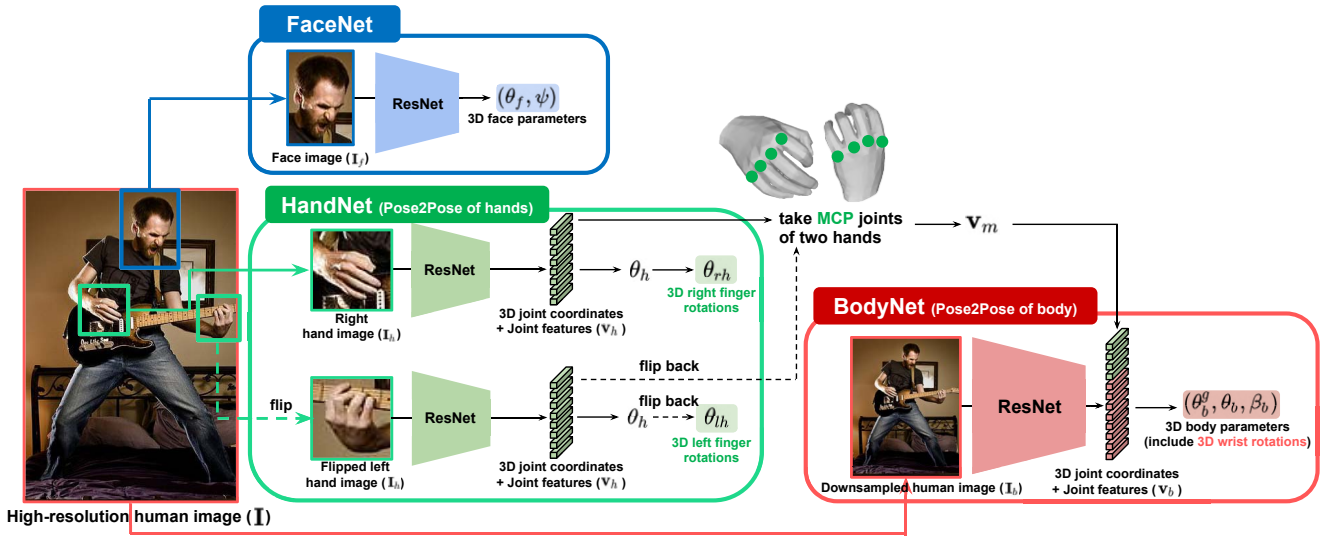

Figure 3. The overall pipeline of Hand4Whole. The BodyNet predicts 3D body parameters, which include 3D wrist rotations, from a combination of body and hand MCP joint information. The HandNet predicts 3D finger rotation only from fine hand information without combining coarse hand information, extracted from the BodyNet. The weights in HandNet for the right and flipped left hand images are shared. The final whole-body 3D human mesh is obtained by forwarding the outputs of BodyNet, HandNet, and FaceNet to the SMPL-X [24] layer. The hand and face boxes are predicted from BodyNet, not shown in the figure. Our Hand4Whole is trained in an end-to-end manner.

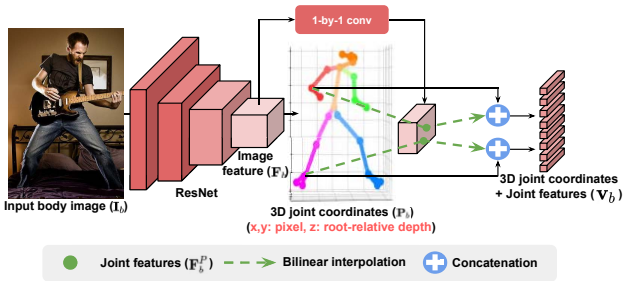

Figure 4. Illustration of the Pose2Pose in our BodyNet. It extracts the joint features by interpolating (x, y) pixel position of the 3D joint coordinates $\mathbf{P}_b$ on the output of 1-by-1 convolution. For the simplicity, we describe only right elbow and ankle.

is passed to a fully-connected layer, which produces the 3D joint rotations θ.

3.2. BodyNet

The BodyNet outputs 3D body joint rotations, SMPL-X shape parameters, camera parameters, and boxes of hands and face. Pose2Pose plays a central role in the prediction of the 3D body joint rotation.

3D body joint rotations. Using the Pose2Pose, BodyNet predicts 3D body joint rotation $\theta_b \in \mathbb{R}^{22\times3}$ from a human image $\mathbf{I}_b \in \mathbb{R}^{3\times256\times192}$, downsampled from a high-resolution human image $\mathbf{I} \in \mathbb{R}^{3\times512\times384}$. The downsampling is necessary to save the computational cost. One modification we made to the Pose2Pose when predicting the 3D joint rotation is that instead of only using $\mathbf{v}_b$, a concatenation of 3D body joint coordinates and body joint features, BodyNet additionally uses $\mathbf{v}_m$, a concatenation of 3D coordinates and features of MCP joints. $\mathbf{v}_m$ is obtained from the HandNet, and Section 3.3 describes how we obtain it in detail. The additional usage of MCP joint information enables BodyNet to produce more accurate 3D wrist rotations. Please note that the predicted 3D body joint rotations θ_b include 3D wrist rotations. The hand MCP joint features are not only beneficial for 3D wrist rotations, but also for 3D elbow rotations, as 3D elbow rotations in the roll-axis are highly related to the hand MCP joints. This is the reason why the hand MCP joint features are used for 3D rotations of all body joints, not restricted to 3D wrist rotations.

SMPL-X shape/camera parameters. The shape parameter $\beta_b \in \mathbb{R}^{10}$ and 3D global translation vector $\mathbf{t}_b \in \mathbb{R}^3$ are predicted from the image feature $\mathbf{F}_b$ using the GAP and a single fully-connected layer.

Hands and face boxes. The BodyNet predicts hand and face bounding boxes by concatenating the image feature $\mathbf{F}_b$ and 3D heatmap $\mathbf{H}_b$ and passing it to two convolutional layers. The 3D heatmap $\mathbf{H}_b$ is predicted for the 3D positional pose of the body joints. The box centers are obtained by applying the soft-argmax [27] to the output of the two convolutional layers. The widths and heights of the boxes are computed by performing bilinear interpolation to $\mathbf{F}_b$ at the box centers and passing the features of each box center to two fully-connected layers.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

3.3. HandNet

Like the BodyNet, the HandNet outputs 3D finger rotations $\theta_h \in \mathbb{R}^{15\times3}$ using Pose2Pose. HandNet takes a hand image $\mathbf{I}_h$, cropped and resized from the high-resolution human image $\mathbf{I}$ by applying RoIAlign [9] at the predicted hand bounding box area. Taking the hand images from the high-resolution image $\mathbf{I}$ instead of the downsampled human image $\mathbf{I}_b$ allows HandNet to utilize detailed finger information. The hand-cropped images of the left hands are flipped to the right hands before being fed to the HandNet. After predicting the 3D finger rotations, we flip back the outputs of the flipped left hands. We denote the 3D finger rotations of the left and right hands as θ_{rh} and θ_{lh}, respectively.

MCP joints features to the BodyNet. As described in Section 3.2, we pass $\mathbf{v}_m \in \mathbb{R}^{3*8+512*8}$, a concatenation of the 3D coordinates and features of eight MCP joints (*i.e.*, four from the right hand and four from the left hand) to the BodyNet for the accurate 3D wrist rotation prediction. To this end, we take the joint features and 3D coordinates of MCP joints of the left and right hands from the outputs of Pose2Pose. The taken 3D coordinates and joint features are concatenated and flattened to a vector $\mathbf{v}_m \in \mathbb{R}^{3*8+512*8}$, which is passed to the BodyNet. Among hand joints, we choose the MCP joints as they show low positional errors than other hand joints while providing essential 3D wrist rotation information.

3.4. FaceNet

We design FaceNet as a simple GAP-based regressor instead of Pose2Pose architecture as deformations of human faces (*e.g.*, facial expressions) are not fully modeled by the 3D joint rotations. The FaceNet predicts 3D jaw rotation $\theta_f \in \mathbb{R}^3$ and facial expression code $\psi \in \mathbb{R}^{10}$ from a face-cropped image $\mathbf{I}_f \in \mathbb{R}^{3\times192\times192}$. The face-cropped image is cropped and resized from the high-resolution human image $\mathbf{I}$ by applying RoIAlign [9] at the predicted face bounding box area to utilize detailed face information. We use ResNet after removing the GAP and fully-connected layer of the last part of the original ResNet.

3.5. Loss functions

Our framework is trained in an end-to-end manner by minimizing the loss function L, defined as follows.

$$L = L_{\text{param}} + L_{\text{coord}} + L_{\text{box}}, \tag{1}$$

where L_{param} is a $L1$ distance between predicted and GT SMPL-X parameters. L_{coord} is a $L1$ distance between predicted and GT joint coordinates, and three types of joint coordinates are used to calculate the loss function: 1) 3D joint coordinates from the Pose2Pose of body and hands, 2) 3D joint coordinates, obtained by multiplying a joint regression matrix of SMPL-X to the 3D mesh, and 3) 2D joint coordinates, obtained by projecting the 3D coordinates from the 3D mesh, to the 2D space using the perspective projection. For the projection, the predicted 3D global translation vector $\mathbf{t}_b$, fixed focal length (5000,5000), and fixed principal points (*i.e.*, a center point of $\mathbf{I}_b$) are used, following [16]. Finally, L_{box} is a $L1$ distance between the predicted and GT center and scale of the hands' and face's boxes.

4. Implementation details

PyTorch [22] is used for implementation. The ResNet of the body branch is initialized with that of Xiao *et al.* [31], pre-trained on MSCOCO 2D human body pose dataset. For the training, we use Adam optimizer [15] with a mini-batch size of 96. Data augmentations, including scaling, rotation, random horizontal flip, and color jittering, are performed during the training. All the 3D rotations are initially predicted in the 6D rotational representation of [33] and converted to the 3D axis-angle rotations. The initial learning rate is set to 10^{-4} and reduced by a factor of 10 at the 10th epoch. A neutral gender SMPL-X model is used for the training and testing. All other details will be available in our codes.

5. Experiment

5.1. Datasets and evaluation metrics

Datasets. For the training, Human3.6M [11], the whole-body version of MSCOCO [12], MPII [1], and FreiHAND [36] are used. The 3D pseudo-GTs for the training are obtained by NeuralAnnot [19]. For the 3D body-only, 3D hand-only, and 3D face-only evaluations, we use 3DPW [29], FreiHAND [36], and Stirling [6], respectively. For the 3D whole-body evaluation, we use EHF [24] and AGORA [23]. We provide qualitative results on the MSCOCO validation set.

Evaluation metrics. Mean per joint position error (MPJPE) and mean per-vertex position error (MPVPE) are used to evaluate 3D joint and mesh vertices positions, respectively. Each calculates the average 3D joint distance (mm) and 3D mesh vertex distance (mm) between the predicted and GT, after aligning a root joint translation. The pelvis is used as the root joint when calculating the 3D errors of the whole body and body. On the other hand, the wrists and neck are used as the root joints when calculating the 3D errors of the hands and face. PA MPJPE and PA MPVPE further align a rotation and scale. The 3D joint coordinates for the MPJPE and PA MPJPE are obtained by multiplying a joint regression matrix, defined in SMPL-X, to the mesh, following previous works [4]. We report average 3D errors of the left and right hands for the 3D hands' error.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

Figure 5. Qualitative comparison of models that doese not take and take hand MCP joint features for the 3D wrist rotation. Utilizing MCP features improves the 3D wrist rotation significantly.

Inputs for 3D wrist prediction	MPVPE (Hands)
Body	50.4
Body + Hand GAP	43.1
Body + All hand joints	43.4
Body + MCP joints (Ours)	**39.8**

Table 1. Comparison of hands MPVPEs, obtained by models that predict 3D wrist rotation from various features, on EHF.

Settings	PA MPVPE (Hands)
With body features	12.3
Without body features (Ours)	**10.8**

Table 2. Comparison of hands PA MPVPEs, obtained by models that predict 3D finger rotations with and without the coarse hand information, on EHF.

Inputs for 3D joint rotations	PA MPVPE (All)
GAP feat.	54.8
Joint feat.	52.2
2D joint coord.	55.8
3D joint coord.	54.3
3D joint coord. + joint feat. (Ours)	**50.3**

Table 3. Comparison of whole-body PA MPVPEs, obtained by models that take various input combinations for the 3D joint rotation prediction, on EHF.

5.2. Ablation study

MCP joint features for 3D wrist rotations. Table 1 and Figure 5 show that Hand4Whole produces accurate 3D wrist rotations by taking a combination of body and hand MCP joint features. The hand MPVPE is measured by separating 3D hand meshes from a whole-body mesh without rotation alignment; therefore, wrong 3D wrist rotation significantly increases the MPVPE of hands. To analyze the effectiveness of the hand MCP joint features, we design three variants. They have exactly the same network architectures as ours except the BodyNet that predicts 3D body joint rotations including 3D wrist rotations. In their BodyNet, they take the same body feature (*i.e.*, a combination of 3D joint coordinates and joint features of the body) and different hand features. Please note that the body feature is essential for plausible 3D wrist rotations when hands are occluded, as shown in Figure 2. As they have the same HandNet architecture, all of them have similar 3D finger rotation predictions. Therefore, most of the hand errors are from wrist rotations, not from finger rotations.

The first variant uses only the body features for the 3D wrist rotation like Zhou *et al*. [34]. It suffers from inaccurate 3D wrist rotation, which indicates additional hand feature is necessary to determine 3D wrist rotation accurately. The second variant uses a combination of the body feature and hand global average pooled (GAPed) feature for 3D wrist rotation prediction like PIXIE [5]. The hand GAP feature is obtained by performing GAP at the output of HandNet's ResNet. It still produces less accurate 3D wrist rotation than ours, which indicates that the hand GAP feature fails to capture essential information of the 3D wrist rotations. The third variant uses a combination of the body features and 3D joint coordinates and joint features of all two hand joints for 3D wrist rotation prediction. It still achieves worse results than ours because the joint features of all two hand joints (*i.e.*, 40 finger joints) contain too much unnecessary information considering that the body consists of 25 joints. From an anatomical point of view, eight MCP joints of two hands mainly contribute to the 3D wrist rotation, as shown in Figure 1, while other finger joints often move independently with the 3D body joint rotation. Our Hand4Whole achieves the lowest 3D errors and accurate 3D wrist rotation by taking body and MCP joint features.

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

Methods	PA MPVPE			MPVPE		
	All	Hands	Face	All	Hands	Face
ExPose [4]	54.5	12.8	5.8	77.1	51.6	35.0
FrankMocap [25]	57.5	12.6	-	107.6	42.8	-
PIXIE [5]	55.0	11.1	**4.6**	89.2	42.8	32.7
Hand4Whole (Ours)	**50.3**	**10.8**	5.8	**76.8**	**39.8**	**26.1**

Table 4. 3D errors comparison on EHF.

Methods	PA MPVPE			MPVPE		
	All	Hands	Face	All	Hands	Face
ExPose [4]	88.0	12.1	4.8	219.8	115.4	103.5
FrankMocap [25]	90.6	11.2	4.9	218.0	95.2	105.4
PIXIE [5]	82.7	12.8	5.4	203.0	89.9	95.4
Hand4Whole (Ours)	**73.2**	**9.7**	**4.7**	**183.9**	**72.8**	**81.6**

Table 5. 3D errors comparison on AGORA.

Removing body features for 3D finger rotations. Table 2 shows that removing the body feature produces better 3D finger rotations. The setting that uses the body feature is similar to previous works [5, 34]. As PA MPVPE is calculated after the rotation alignment, aligned outputs have almost correct 3D wrist rotations. Therefore, most of the 3D errors come from 3D fingers. To analyze how the body feature affects the 3D finger rotation, we design a variant whose HandNet takes additional coarse hand information, obtainable from hand areas at the body feature. To extract the coarse hand information, we perform RoIAlign [9] to the output of the first residual block in the BodyNet's ResNet using the predicted hand bounding box. Then, the output of RoIAlign is added in an element-wise manner to the first residual block in the HandNet's ResNet. The reason why only removing the body feature produces better 3D finger rotation is that the coarse hand information in the body feature contains various unnecessary information, such as body and backgrounds, while barely having finger information due to the small resolution of hands. As such unnecessary information corrupts the hand feature of the HandNet, using only the hand feature from the HandNet like 3D hand-only reconstruction methods [2, 17, 35], produces better 3D finger rotations.

Effectiveness of joint features. Table 3 shows that the joint feature of our Pose2Pose is much more beneficial than previous GAPed feature [4, 5, 25]. In addition, combining the joint feature with 3D joint coordinates like ours achieves the best results. For the demonstration, we designed four variants that take different combinations of features for the 3D body and hand joint rotation prediction. The first and second variants take a GAP feature vector and joint feature, respectively. The GAP feature is obtained by performing GAP to the output of ResNet. As GAP marginalizes the spatial domain, the feature vector losses detailed human articulation information. On the other hand, our joint feature preserves such detailed articulation information by interpolating human joint positions, beneficial for accurate 3D joint rotation prediction. The third and fourth variants take 2D and 3D joint coordinates, respectively. Compared to the third variant, the fourth variant achieves lower PA MPVPE by utilizing additional depth information. Finally, we combine the 3D joint coordinates and joint features for the best results.

Methods	MPJPE	PA MPJPE
*** Body-only methods**		
HMR [14]	130.0	81.3
SPIN [16]	96.9	59.2
Pose2Mesh [3]	**88.9**	58.3
I2L-MeshNet [20]	93.2	57.7
ROMP [28]	91.3	**54.9**
*** Whole-body methods**		
ExPose [4]	93.4	60.7
FrankMocap [25]	-	61.9
PIXIE [5]	-	61.3
Hand4Whole (Ours)	**86.6**	**54.4**

Table 6. 3D body error comparison on 3DPW.

Methods	PA errors	F scores
*** Hand-only methods**		
FreiHAND [36]	10.7 / -	0.529 / 0.935
Pose2Mesh [3]	7.8 / 7.7	0.674 / 0.969
I2L-MeshNet [20]	7.6 / 7.4	0.681 / 0.973
METRO [18]	**6.7 / 6.8**	**0.717 / 0.981**
*** Whole-body methods**		
ExPose [4]	11.8 / 12.2	0.484 / 0.918
Zhou *et al.* [34]	- / 15.7	- / -
FrankMocap [25]	11.6 / 9.2	0.553 / 0.951
PIXIE [5]	12.1 / 12.0	0.468 / 0.919
Hand4Whole (ResNet-18)	8.6 / 8.6	0.621 / 0.962
Hand4Whole (Ours)	**7.7 / 7.7**	**0.664 / 0.971**

Table 7. 3D hand errors (PA MPVPE/PA MPJPE and F-score@5mm/15mm) comparison on FreiHAND.

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

EHF (Whole-body evaluation benchmark). Table 4 shows that our Hand4Whole largely outperforms all previous whole-body methods on EHF. As existing works mainly have reported PA metrics, we use their released codes and pre-trained weights to report MPVPEs.

AGORA (Whole-body evaluation benchmark). Table 5 shows that our Hand4Whole largely outperforms all previous whole-body methods on AGORA. We obtained all numbers using their released codes with pre-trained weights and an official evaluation tool[2]. We use the same human bounding boxes for all methods.

3DPW (Body-only evaluation benchmark). Table 6 shows that our Hand4Whole largely outperforms all previ-

[2]https://github.com/pixelite1201/agora_evaluation

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

Figure 6. Qualitative comparison of the proposed Hand4Whole, ExPose [4], FrankMocap [25], and PIXIE [5]. Implausible 3D wrist rotations are highlighted.

ous whole-body methods on 3DPW, although it is slightly beaten by recent body-only methods. The stronger performance on the 3D body benchmark than previous whole-body methods results from a combination of 3D joint coordinates and joint features for the 3D body joint rotation prediction. On the other hand, all previous whole-body methods use GAP features for the 3D body joint rotation prediction.

FreiHAND (Hand-only evaluation benchmark). Table 7 shows that our Hand4Whole largely outperforms all previous whole-body methods on FreiHAND, although it is slightly beaten by recent hand-only methods. We obtained the results of FrankMocap using their released codes and pre-trained weights as they did not report their results on FreiHAND. In particular, Hand4Whole achieves the best results even after changing our hand branch ResNet-50 to ResNet-18 following ExPose [4].

MSCOCO (Qualitative comparison of the whole-body). Figure 6 shows that our Hand4Whole produces much more accurate 3D wrists and finger rotations than previous whole-body methods.

Overall, Hand4Whole largely outperforms all previous whole-body methods. In particular, it produces much better 3D hand results. The comparisons are consistent with the results of the ablation studies.

6. Conclusion

We present Hand4Whole, a whole-body 3D human mesh estimation system that produces much better 3D hands in the whole-body 3D mesh. Hand4Whole utilizes hand MCP joint features for the 3D wrist rotation by introducing Pose2Pose. In addition, it discards the body feature when predicting 3D finger rotations. Hand4Whole largely outperforms previous 3D whole-body human mesh estimation methods on all benchmarks.

Acknowledgements. This work was supported in part by IITP grant funded by the Korea government (MSIT) [No. 2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 3, 7

[3] 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, 7

[4] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. Monocular expressive body regression through body-driven attention. In *ECCV*, 2020. 1, 2, 3, 5, 7, 8

[5] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. Collaborative regression of expressive bodies using moderation. In *3DV*, 2021. 1, 2, 3, 6, 7, 8

[6] Zhen-Hua Feng, Patrik Huber, Josef Kittler, Peter Hancock, Xiao-Jun Wu, Qijun Zhao, Paul Koppen, and Matthias Rätsch. Evaluation of dense 3D reconstruction from 2D face images in the wild. *FG*, 2018. 5

[7] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3D hand shape and pose estimation from a single RGB image. In *CVPR*, 2019. 3

[8] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 3

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, 2017. 5, 7

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[11] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 2014. 5

[12] 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. 1, 5

[13] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3D deformation model for tracking faces, hands, and bodies. In *CVPR*, 2018. 3

[14] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3, 7

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014. 5

[16] 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. 3, 5, 7

[17] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 3, 7

[18] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 3, 7

[19] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. NeuralAnnot: Neural annotator for 3D human mesh training sets. In *CVPRW*, 2022. 5

[20] 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. 3, 7

[21] Lea Muller, Ahmed AA Osman, Siyu Tang, Chun-Hao P Huang, and Michael J Black. On self-contact and human pose. In *CVPR*, 2021. 3

[22] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 5

[23] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in geography optimized for regression analysis. In *CVPR*, 2021. 5

[24] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, 2019. 1, 3, 4, 5

[25] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. FrankMocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshop*, 2021. 1, 2, 3, 7, 8

[26] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *CVPR*, 2019. 3

[27] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 3, 4

[28] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3d people. In *ICCV*, 2021. 3, 7

[29] Timo von Marcard, Roberto Henschel, Michael J Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *ECCV*, 2018. 5

[30] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. Monocular Total Capture: Posing face, body, and hands in the wild. In *CVPR*, 2019. 3

[31] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 5

[32] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3D human shape and articulated pose models. In *CVPR*, 2020. 3

[33] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 5

[34] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. Monocular real-time full body capture with inter-part correlations. In *CVPR*, 2021. 1, 2, 3, 6, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.

[35] 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, 7

[36] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. FreiHAND: A dataset for markerless capture of hand pose and shape from single RGB images. In *ICCV*, 2019. 5, 7

Authorized licensed use limited to: ShanghaiTech University. Downloaded on October 19,2026 at 16:05:11 UTC from IEEE Xplore. Restrictions apply.