Monocular Expressive 3D Human Reconstruction of Multiple People

Zhenghao Zhao
Illinois Institute of Technology
Chicago, IL, USA
zzhao48@hawk.iit.edu

Hao Tang
Carnegie Mellon University
Pittsburgh, PA, USA
bjdxtanghao@gmail.com

Joy Wan
University of Illinois Urbana-Champaign
Urbana-Champaign, IL, USA
joywan2@illinois.edu

Yan Yan
Illinois Institute of Technology
Chicago, IL, USA
yyan34@iit.edu

ABSTRACT

Whole-body pose estimation aims to regress human pose models that include the body, hand, and facial details from RGB images. While the task of whole-body mesh recovery has been extensively studied in recent literature, the focus has predominantly been on human mesh recovery for a single person, despite the frequent occurrence of multiple people in practical scenarios. Similar to body-only cases, such single-person whole-body pose estimation methods often fail in the multiple-people problem for two reasons: (i) Given the ambiguous bounding box, which could contain more than one instance, it is difficult for single-person-oriented methods to regress the body mesh model of the target person. (ii) Single-person pose estimation approaches neglect the person-person occlusions and the depth order among instances, thus generating interpenetrated models. In this paper, we propose the **M**ulti-person **E**xpressive **PO**se (MEPO) model, which exploits expressive 3D human model reconstruction for multiple people. To our best knowledge, our model is the first multi-person whole-body mesh reconstruction model, which is intensified by heatmap, depthmap, and depth order loss. We propose the **H**eatmap **E**nhancement **Net** (HENet) to leverage the heatmap information to assist the model in concentrating on the target person in crowded multi-person cases, while the depthmap delivers depth information of the image. Furthermore, we impose a depth order loss to recover human mesh precisely for overlapped people. In our experiments, we evaluate our model on multiple challenging datasets, including AGORA, which consists of complex occlusions similar to real-world scenarios. Our method has a significant performance improvement compared with the state-of-the-art pose estimation methods.

CCS CONCEPTS

• **Computing methodologies** → **Reconstruction**.

ICMR '24, June 10–14, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0619-6/24/06
https://doi.org/10.1145/3652583.3658092

KEYWORDS

3D Pose Estimation; Whole-body Pose Estimation; Multi-person Pose Estimation

ACM Reference Format:
Zhenghao Zhao, Hao Tang, Joy Wan, and Yan Yan. 2024. Monocular Expressive 3D Human Reconstruction of Multiple People. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand*. Phuket, Thailand, 10 pages. https://doi.org/10.1145/3652583.3658092

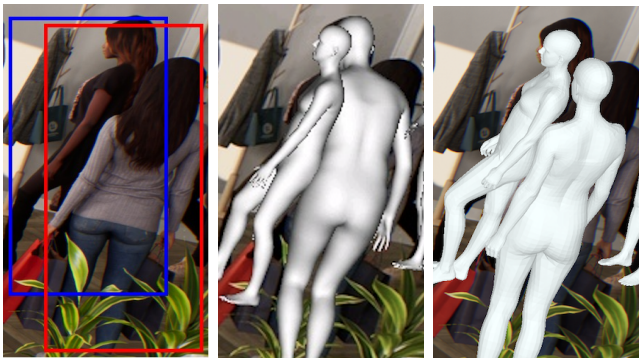
1 INTRODUCTION

With the grooming of deep learning [44], 3D human pose estimation has been drawing broader attention recently. It is involved in various applications, such as sign language understanding [40], AR [7], VR [5], robotics [27], and human-object interaction detection [51]. To achieve a more accurate deformation of the human pose, methods of reconstructing a parametric human model from a single image have been introduced [25, 28, 56] in the past few years. Given a single RGB image, these methods estimate human body joints and recover a parametric human model, such as the widely used SMPL [33] model from the image.

In recent years, the academic community has expanded its focus beyond body-only pose estimation to include the intricacies of hand, finger, and facial movements. These elements are crucial for understanding human activities in a more comprehensive manner. Advanced techniques have been developed to reconstruct detailed models of human hands [16, 26, 29, 47], faces [11, 15, 48, 49], and entire human bodies [10, 14, 35, 38, 43].

Whole-body methodologies such demonstrate the ability to recover whole-body models from keypoint detection. Other approaches, like FrankMocap [43], ExPose [10], PIXIE [14], and Pose2Pose [35], adopt a segmented strategy. This involves initially dividing the subject into parts such as the head, body, and hands. Each part is then processed to regress its corresponding model. These models are subsequently integrated to form a cohesive and expressive human model. This segmented approach allows for a more detailed and accurate representation of the human form, enhancing the understanding of human body dynamics in various contexts.

However, such single-person methods predominantly employ a top-down approach. This involves initially identifying the individual within an image, followed by the reconstruction of the pose mesh within the defined bounding box. However, applying these



(a) Detection result (b) Baseline (c) MEPO (Ours)

Figure 1: Human reconstruction in the crowded scene. In Figure 1a, the detection bounding box (red) of the woman in the front is ambiguous. Existing whole-body pose estimation method [43] cannot handle such ambiguity and depth order relation between overlapped people (Figure 1b), while our model can reconstruct human models correctly under this crowded situation (Figure 1c).

single-person methodologies to multi-person pose estimation tasks presents two primary challenges. Firstly as pointed out in [45], in scenarios involving multiple people, a single image crop might contain more than one instance, and single-person pose estimation methods fail to estimate body pose for such an ambiguous image crop. For example, as depicted in Figure 1a, the bounding box surrounding the foremost woman also includes a considerable portion of the second woman's body. In Figure 1b, due to the ambiguous bounding box, the pose estimation model incorrectly recognizes the head of the woman behind as the head of the woman in the front. Secondly, these single-person-oriented approaches ignore the depth order of overlapping people, which leads to inconsistent depth ordering and mesh model interpenetration. As shown in Figure 1b, the body meshes of two people are interpenetrated because the model simply recovers their body mesh models but is unaware of the depth order relations of two instances.

Expanding our perspective, it becomes apparent that current methods for reconstructing whole-body human mesh models scarcely address the complexities of multi-person pose estimation. To the best of our knowledge, the majority of existing whole-body pose estimation techniques are designed with a single-person focus. Conversely, most current multi-person pose estimation methods [9, 20, 45, 57] primarily target body-only mesh regression, leaving the area of multi-person whole-body pose estimation relatively uncharted.

To address the aforementioned challenges, we propose a multi-person whole-body pose estimation model, **M**ulti-person **E**xpressive **PO**se (MEPO). The method consists of two stages. The first stage of the model detects instances from the raw image, and the second stage reconstructs the whole human body model for each detected instance. Specifically, we propose a novel **H**eatmap **E**nhancement **Net** (HENet) to address the ambiguous bounding box issue mentioned previously. We first generate a heatmap that represents the existence information of each instance, and then we apply the heatmap to the intermediate image features to assist the network in recognizing each instance. This way, the model is designed to focus on the target person even in ambiguous bounding boxes and regress the human model correspondingly. Meanwhile, to address the depth order issue on the overlapping people, we introduce the depthmap and the depth order loss into our model to handle the depth order relations among instances. We evaluate our model on multiple datasets [37, 39, 50], and our model outperforms state-of-the-art methods. As shown in Figure 1c, our model is robust to crowded scenes which contain overlapped instances.

The contributions of our paper are summarized as follows: **(i)** We propose a novel multi-person pose estimation method named MEPO. To the best of our knowledge, our model is the first multi-person pose estimation method to reconstruct expressive human parametric models with finger details and facial expressions. **(ii)** We introduce a novel HENet to alleviate the ambiguity of the implicit bounding box to improve the mesh reconstruction model performance in crowded scenes. Moreover, we introduce depthmap and depth order loss to handle the depth order of overlapping instances. **(iii)** We evaluate our model on multiple datasets, including a challenging dataset, AGORA [37], and our proposed model outperforms state-of-the-art whole-body pose estimation methods.

2 RELATED WORK

2.1 Human Modeling

The simplest way to represent human poses is to use 2D/3D key points to represent body joints. However, keypoints based methods can not provide sufficient information for human behavior analysis. Instead of using 2D or 3D keypoints, 3D parametric human models describe human motions vividly and accurately. The main idea of 3D parametric human models is to model the deformation of 3D humans by a series of parameters. With a 3D parametric human model, we can describe the human pose more accurately than just using keypoints. The first successful approach in this field is SCAPE [2]. The most commonly used model in academia, SMPL [33] is a skinned vertex-based model, which contains 6890 vertices and 23 joints.

Similarly to the human body model, models for hand deformation, such as MANO [42] and models for face deformation, such as 3DMM [3] and FLAME [30], have also been proposed. In addition, recently, researchers proposed parametric human models that not only represent a part of the human body, but also model the human body together with hands or face. SMPLH [42] combines the SMPL model with the MANO hand model to capture body and hands motions together. SMPL-X [39] model extends the SMPL body model with the MANO hand model and the FLAME head model. SMPL-X model contains 10475 vertices and 54 joints, with which 24 joints for body representation and the rest for hands and face, including jaw, eyeballs, and fingers.

In our work, we follow the recent whole-body pose estimation works [10, 14, 35, 38] and take the SMPL-X model as our output format.

2.2 Multi-Person Pose Estimation

Monocular multi-person pose estimation is to predict joints or regress human mesh models of all the people from a single RGB image. Handling occlusions and determining depth orders among individuals are crucial aspects to consider in multi-person pose

estimation methods of this nature. According to the different design strategies used in the model, multi-person body mesh recovery methods can be categorized into two types: two-stage methods and one-stage methods. Two-stage methods are straightforward methods in multi-person pose estimation, the main idea of which is to convert multi-person tasks to single-person pose estimation tasks by first detecting and cropping instances from the image with the off-the-shelf detectors [17] and applying body model recovery methods for each instance. However, truncations, occlusions, and interpenetrations in multi-person cases drive it hard for the target person to regress correctly. Therefore, handling person-person occlusions and depth orders among instances becomes important in top-down approaches. Jiang *et al.* [20] introduces the interpenetration loss to avoid colliding models and the pixel level depth ordering-aware loss based on instance segmentation to reconstruct more accurate human models. 3DCrowdNet [9] generates a 2D pose heatmap for the target person based on the off-the-shelf 2D pose estimator and passes it to a joint-based body mesh regressor. Recently, some papers have tried to find one-stage solutions for multi-person pose estimation. One-stage methods not only execute faster but also omit the ambiguous bounding box shortage of top-down methods. For example, Sun *et al.* [45] propose ROMP to handle the occlusions and inference in real-time. Zhang *et al.* [57] propose a one-stage solution, which adopts an FPN [24] to represent the depth level for each instance and applies inter-instance ordinal relation supervision to handle occlusions.

Despite the success of various multi-person pose estimation methods, few explore the possibility of whole-body pose estimation in multi-person cases. Unlike existing whole-body multi-person pose prediction methods, such as [6] and [53], our model reconstructs a 3D human mesh model instead of just predicting the human pose as keypoints.

2.3 Whole-Body Pose Estimation

Whole-body (full-body) pose estimation aims to recover not only human body models but also expressive models with hands and faces. At an early time, whole-body pose estimation is a challenging task because of the lack of datasets, the absence of the human parametric model for whole-body representation, and the inconsistent annotations from various datasets. Consequently, researchers have to train each part of the model separately on different datasets.

Besides the development of off-the-shelf methods [15, 26] to reconstruct the expressive face and hand models, datasets [37, 38, 55, 58] with SMPL-X [39] format annotations published in recent years also accelerate the progress on whole-body pose estimation, in which EHF [38], THUman2.0 [55], and MultiHuman [58] are small datasets that contain hundreds of images for evaluation. SMPL-X [39], the model developed from SMPL by adding MANO [41] as the hand model and FLAME [31] as the face model, has recently received increasing attention. Recent single-person pose estimation approaches [10, 38, 43, 59] adopt SMPL-X model. For instance, ExPose [10] introduces body-driven attention to extract high-resolution image crops for hands and face to improve performance. AGORA [37] is a recently released large synthetic dataset with 15,754 images with ground-truth SMPL-X fittings.

Optimization-based approaches such as SMPLify-X [39] follow a similar procedure of SMPLify [4] by taking detected 2D keypoints and optimizing the model parameters to reconstruct the model. Most regression-based works [10, 38, 43, 59, 59] first regress model parts on the body, hands, and face separately and then combine all parts together. For example, Zhou *et al.* [59] recovers the face with the 3DMM [12] face model and adopts SMPLH-neutral as the body-hand joint model. FrankMocap [43] utilizes an integration module to combine face, body, and hand parts into a whole-body representation.

However, to the best of our knowledge, a whole-body method for multi-person pose estimation remains absent in this area.

3 METHODOLOGY

In this paper, we introduce a novel approach for multi-person whole-body mesh reconstruction, which tackles the challenges outlined in Section 1. Specifically, our model addresses the issue of ambiguous bounding boxes by utilizing a Heatmap Enhancement Net (HENet), which effectively leverages heatmap information for localizing the target instance in the bounding box. Additionally, to resolve the complexity of determining the depth order of overlapping instances, we have integrated a depthmap and a depth order loss mechanism into our model. In this section, we begin by presenting a structured overview of our model framework, followed by comprehensive descriptions of its essential components.

3.1 Overall Framework

Given an image I, we propose a method that seeks to detect all the person instances $\mathbf{P}$ in the image I and regresses an SMPL-X [39] mesh model $\mathbf{M}_{p_i}$ including body, hands, and face for each person instance $p_i \in \mathbf{P}$. Our approach is a two-stage pose estimation method, where the first stage detects instances from the image, and the second stage reconstructs human mesh models for detected instances. Furthermore, we propose to use a HENet to deal with the ambiguous bounding box problem and introduce the depthmap and the depth order loss to handle person-person occlusions in multi-person scenarios.

Figure 2 shows an overview of the model architecture. We first use off-the-shelf 2D pose estimation methods to detect instances p_i from the image I and output their corresponding 2D joints $\mathbf{J}_{p_i}$. In the meantime, we use a DepthNet to output the image depthmap $\mathbf{F}_D$. We pass each detected instance into FaceBranch, HandBranch, and BodyBranch. Following the approach of [34], for FaceBranch, we process the face crop with ResNet [18], followed by a fully connected layer to get face expression $\psi \in \mathbb{R}^{10}$ and jaw joint $\theta_f \in \mathbb{R}^3$. For HandBranch, we first apply ResNet on hand crops to get hand features $\mathbf{F}_{hand}$, and then we employ the PoseNet to $\mathbf{F}_{hand}$ to get hand joint feature $\mathbf{F}_{J_{hand}}$, which is concluded as:

$$\mathbf{F}_{J_{hand}} = f_{PoseNet}(f_{ResNet}(I_{hand})). \tag{1}$$

A fully connected layer is applied to the hand joint feature $\mathbf{F}_{J_{hand}}$ to get finger joints $\theta_h \in \mathbb{R}^{45}$. The hand joint feature will also be concatenated to the body joint feature $\mathbf{F}_{J_{body}}$.

In the BodyBranch, we process the image crop with convolutional layers to get the intermediate image feature $\mathbf{F}_{p_i}$. The HENet takes heatmap, and $\mathbf{F}_{p_i}$ as inputs and produces heatmap-enhanced image feature $\mathbf{F}_{he}$. $\mathbf{F}_{he}$ is then concatenated with the depthmap $\mathbf{F}_D$,

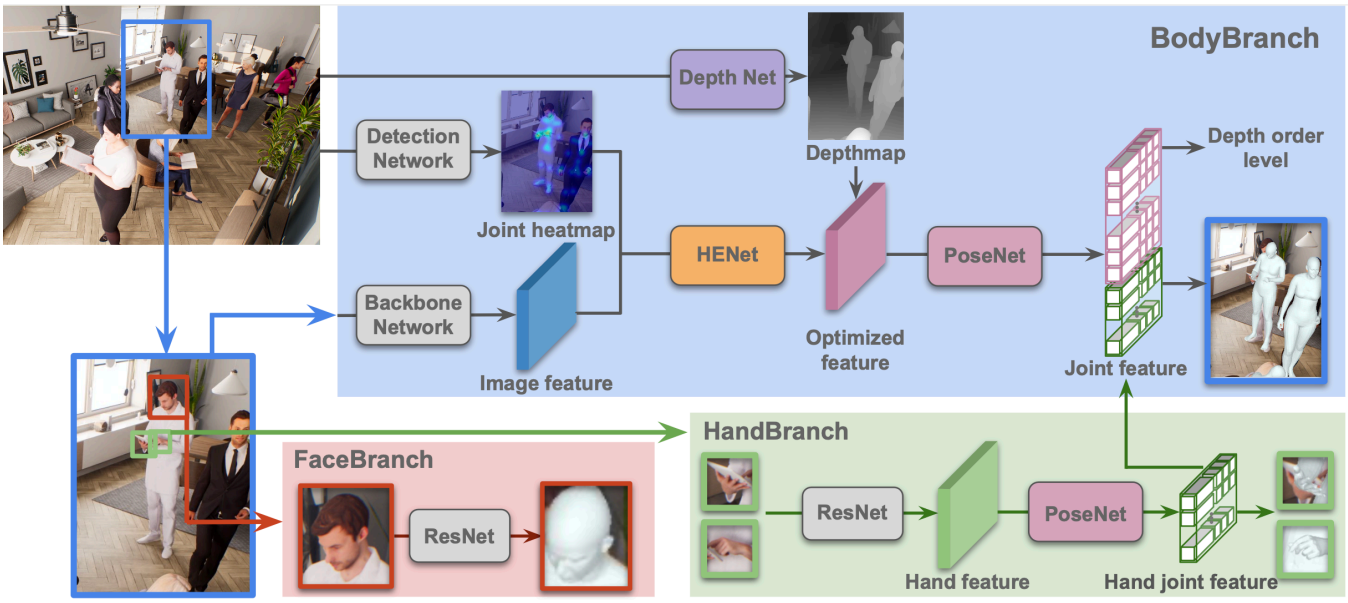


Figure 2: Overview of the proposed MEPO. We use 2D pose estimation model to generate instance proposals as well as their 2D joint heatmaps from the original image. For each detected instance, we pass it to BodyBranch, FaceBranch, and HandBranch to get an expressive human model. We use the HENet to emphasize the existence of the target person in the image crop. Additionally, the DepthNet outputs the depthmap from the original image, which is combined with heatmap enhanced image feature to get the depth-aware optimized feature. We pass the optimized image feature to PoseNet to recover the SMPL-X [39] mesh model. Furthermore, the model outputs a depth order level for each instance, and the depth order loss is applied to address the ambiguous depth order issue among overlapping instances.

which is concluded as:

$$\mathbf{F}_{p_0}, \mathbf{F}_{p_1}, ..., \mathbf{F}_{p_n} = f_{CNN}(f_{openpose}(I)), \tag{2}$$

$$\mathbf{F}'_i = f_{HENet}(\mathbf{F}_{p_i}, \mathbf{F}_{h_i}) \oplus \mathbf{F}_{d_i}, \tag{3}$$

where $\mathbf{F}_{h_i}$ is the heatmap feature of the instance p_i, $\mathbf{F}'_i$ is the depth-aware optimized image feature, and $\mathbf{F}_{d_i}$ is the corresponding depthmap of p_i from $\mathbf{F}_D$. Then we pass $\mathbf{F}'_i$ to the PoseNet to get the body joint feature $\mathbf{F}_{J_{body}}$, and concatenate it with the hand joint feature $\mathbf{F}_{J_{hand}}$ from the HandBranch to output body pose $\theta_b \in \mathbb{R}^{63}$, body shape $\beta \in \mathbb{R}^{10}$ and camera parameters $\mathbf{K} \in \mathbb{R}^3$.

3.2 HENet

As discussed in Section 1, what we are trying to solve is not only a whole-body pose estimation problem but also a multi-person pose estimation problem. For multi-person pose estimation, especially in crowded scenes, the instance detection result at the first stage is likely to contain more than one instance. The existing whole-body human reconstruction model performs poorly due to such ambiguous detection results. Therefore, we propose to leverage the heatmap information to impose the reconstruction model focus on the target person in an ambiguous bounding box. Though for multi-person pose estimation methods, there are several intends [9, 23, 45] to take advantage of the heatmap, our proposed method utilizes the heatmap information more efficiently. We will begin by outlining the advantages of our design over the aforementioned methods in this section and then illustrate the performance gap between our method and these methods in Section 4.3 and Section 4.4.

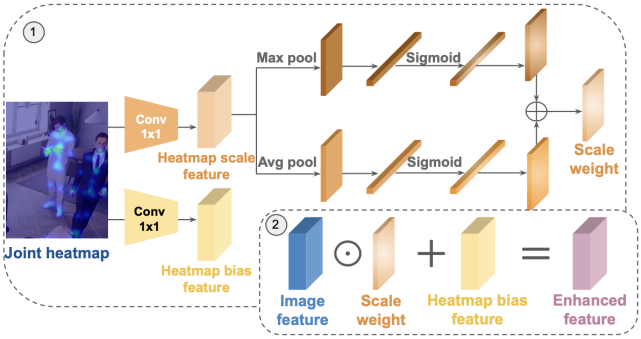


Figure 3: Structure of HENet in our model. Inside HENet, we extract the heatmap scale feature and heatmap bias feature from the heatmap and apply them to the intermediate image feature.

For each instance p_i in the image I, we first predict the center joints $\mathbf{J}_{p_i} \in \mathbb{R}^{Jnum \times 2}$ of the person p_i, where $Jnum$ denotes the number of joints. With predicted joint coordinates, we apply a 2D Gaussian distribution to each selected joint. Next, to assist the mesh reconstruction model in focusing on the target instance, we process intermediate image features as well as their corresponding heatmap in the **H**eatmap **E**nhancement **Net** (HENet). The design of HENet

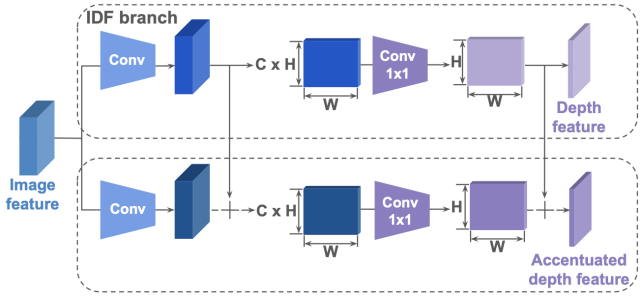


Figure 4: Structure of DepthNet in our model. We employ two Image Depth Feature (IDF) branches to get the depth feature and accentuated depth feature, which is concatenated with the heatmap-enhanced feature.

aims to make the model focus on the target person in the image. To this end, we adopted the concept of spatial attention [52, 54] and utilized it to generate the scale weight feature in the scale feature branch of the HENet.

The structure of our proposed HENet is shown in Figure 3. For an intermediate image feature $\mathbf{F}_{p_i} \in \mathbb{R}^{C\times W\times H}$, we apply convolutional layers on the heatmap and obtain heatmap scale feature and bias feature with the same size of the image feature independently, $\mathbf{F}_{hs} \in \mathbb{R}^{C\times W\times H}$ and $\mathbf{F}_{hb} \in \mathbb{R}^{C\times W\times H}$ respectively, which are processed along with the image feature $\mathbf{F}_{p_i}$ through HENet. With the obtained heatmap scale feature $\mathbf{F}_{hs} \in \mathbb{R}^{C\times W\times H}$, for each pixel on the scale feature, we apply the max pooling and average pooling along the channels. We can obtain max pool $\mathbf{F}_{max} \in \mathbb{R}^{1\times W\times H}$ and average pool $\mathbf{F}_{avg} \in \mathbb{R}^{1\times W\times H}$, respectively. Then we reshape $\mathbf{F}_{max}$ and $\mathbf{F}_{avg}$ into $\mathbb{R}^{1\times(W\times H)}$, and apply a sigmoid function. Then we reshape them back to $\mathbb{R}^{1\times W\times H}$, and add two features element-wisely to obtain the scale weight $\mathbf{F}_w \in \mathbb{R}^{1\times W\times H}$. The scale weight $\mathbf{F}_w$ is multiplied element-wisely to the image feature, and the heatmap bias feature $\mathbf{F}_{hb}$ is added element-wisely to the image feature, which is formulated as:

$$\mathbf{F}_{he} = \mathbf{F}_{p_i} \odot \mathbf{F}_w + \mathbf{F}_{hb}, \tag{4}$$

where $\mathbf{F}_{he}$ is the enhanced image feature from the HENet.

3.3 DepthNet and Depth Order Loss

Besides ambiguous bounding boxes, the depth order relation between instances in the image is also an unsolved problem for existing whole-body pose estimation methods. In our method, we propose to leverage depth information to deal with depth order among instances. Specifically, we propose to use a DepthNet to capture depth information from the image and use a depth order loss to handle the depth order relations.

DepthNet: To obtain the depth information from the image, we built a DepthNet in our model. The structure of the DepthNet is shown in Figure 4. Given the image feature $\mathbf{F}_I \in R^{C\times H\times W}$, we use two IDF branches to get the depth feature $\mathbf{F}_d$ and the accentuated depth feature $\mathbf{F}_{ad}$ respectively. Inside each **I**mage **D**epth **F**eature (IDF) branch, we apply a 1-by-1 convolution to $\mathbf{F}_I$ to change its channel. Then we reshape the output feature $\mathbf{F}'_I$ to $\mathbb{R}^{(C\times H)\times W}$. After reshaping, we employ three 1-by-1 convolutional layers to change the feature size from $\mathbb{R}^{(C\times H)\times W}$ to $\mathbb{R}^{W\times H}$ and obtain the output depth feature $\mathbf{F}_d \in \mathbb{R}^{1\times W\times H}$. For the branch to output accentuated depth feature $\mathbf{F}_{ad}$, we additionally sum the intermediate feature $\mathbf{F}'_I$ and the reshaped feature $\mathbf{F}_d$ from the Depth feature branch to the accentuated features. The depthmap $\mathbf{F}_D$ is obtained by concatenating the depth feature $\mathbf{F}_d$ and the accentuated depth feature $\mathbf{F}_{ad}$ together. After we get the depthmap, we concatenate the corresponding depthmap $\mathbf{F}_{d_i}$ with the enhanced heatmap $\mathbf{F}_{he}$ to get the depth-aware optimized image feature.

Algorithm 1 Ground-Truth Depth Order Extraction.

Initialization
d_i is the depth level of instance p_i
$\mathbf{D}=\{d_0, d_1, ..., d_n\}=\{0, 0, ..., 0\}$
for each instance p_i **do**
  if p_i is occluded by p_j **then**
    if $d_i < d_j$ **then**
      # O_i is the set of occluders of p_i
      Add p_j to O_i
      UPDATE_DEPTH(p_i, O_i)
procedure UPDATE_DEPTH(p_i, O_i)
  for $p_k \in O_i$ **do**
    if $d_k \leq d_i$ **then**
      $d_k = d_i + 1$
      UpdateDepth(p_k, O_k)

Depth Order Loss: As mentioned in Section 1, single-person pose estimation methods are unaware of the depth order of overlapping instances in the image, leading to inconsistent depth ordering and mesh model interpenetration. Therefore, we propose to use depth order level d_i for each instance p_i to indicate the depth order relations among people in the image. A person who occludes others has a higher depth order level. Jiang *et al.* [20] proposed a depth order loss pixel-wisely, while we found that compared to the pixel-wisely loss, depth order loss at the instance level is more stable. Inspired by [8], we propose an instance-level depth order loss as illustrated below.

As shown in Figure 2, after obtaining body joint features $\mathbf{F}_{J_{body}}$, we output a depth order level $\hat{d}_i$ of the instance. We can further define the depth order loss as

$$L_{(p_i)} = \begin{cases} \log(1+\exp(|d_i - \hat{d}_i|)), & |d_i - \hat{d}_i| \leq T, \\ (d_i - \hat{d}_i)^2, & |d_i - \hat{d}_i| > T, \end{cases} \tag{5}$$

where d_i is the ground-truth depth order level, and T is a predefined threshold to determine whether the predicted $\hat{d}_i$ and d_i are on the same depth order level. Then we can compute the total depth order loss for the entire image by averaging the loss of all the instances in the image, which is

$$L_{depth} = \frac{1}{N}\sum_{p_i\in\mathbf{P}} L_{(p_i)}, \tag{6}$$

where N denotes the number of detected instances in the image, and $\mathbf{P}$ denotes all detected instances.

The ground-truth depth order levels in our experiments are defined and obtained from ground-truth mask annotations as shown in Algorithm 1. The basic idea of the algorithm is that if an instance occludes others, we update its depth order level as well as its occluders' depth order level.

Method	MPJPE ↓				MVE ↓				NMJE ↓		NMVE ↓	
	B	LH/RH	F	FB	B	LH/RH	F	FB	B	FB	B	FB
SMPLify-X [39]	182.1	46.5/49.6	52.9	231.8	187.0	**48.3**/51.4	48.9	236.5	256.5	326.5	263.3	333.1
ExPose [10]	150.4	72.5/68.8	55.2	215.9	151.5	74.9/71.3	51.1	217.3	183.4	263.3	184.8	265.0
Frankmocap [43]	165.2	52.3/53.1	-	-	168.3	54.7/55.7	-	-	204.0	-	207.8	-
PIXIE [13]	140.3	**46.4**/**46.0**	54.5	189.3	142.2	49.5/**49.0**	50.2	191.8	171.1	230.9	173.4	233.9
Hand4Whole [34]	87.1	47.5/48.6	52.1	136.5	91.1	49.5/50.5	48.5	140.6	99.2	155.5	103.8	160.2
MEPO (Ours)	**85.8**	47.3/48.3	**51.9**	**134.9**	**89.1**	49.0/50.2	**48.3**	**138.3**	**97.1**	**153.2**	**102.2**	**158.6**

Table 1: Quantitative comparisons with the SOTA whole-body methods. We compare our model with existing whole-body methods on the AGORA dataset with whole-body settings. B, LH, RH, F, and FB denote corresponding metrics for body, left hand, right hand, face, and whole-body, respectively.

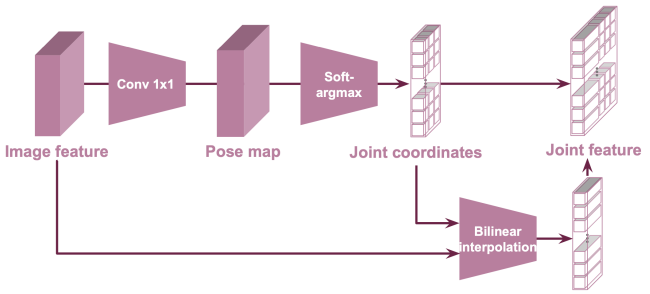


Figure 5: Structure of the PoseNet in our model. We first use the soft-argmax function to calculate joint coordinates via Pose map. Then we use bilinear interpolation to sample the joint features from the image feature.

3.4 PoseNet

We use PoseNet to obtain SMPL-X parameters from the image features. The structure of the PoseNet is shown in Figure 5. Given the input image feature $\mathbf{F}'_i$, a 1-by-1 convolutional layer is applied and outputs a pose map $\mathbf{P}_i$. Then we use the soft-argmax to compute the coordinates from pose map $\mathbf{P}_i$. For each coordinate we get, we apply the bilinear interpolation on the input image feature $\mathbf{F}'_i$ to obtain the features for each joint. Then we concatenate the obtained features and joint coordinates together to get the joint feature $\mathbf{F}_{J_i}$. The joint feature $\mathbf{F}_{J_i}$ is passed to a fully connected layer to obtain the pose parameters.

The loss function for our approach is formulated as:

$$L = L_{pose} + L_{depth}, \tag{7}$$

where L_{depth} is the depth order loss introduced in Section 3.3. L_{pose} is the total pose estimation loss, which is defined as:

$$L_{pose} = L_{mesh} + L_{joint} + L_{bbox}, \tag{8}$$

where L_{mesh} computes the L_1 loss between the predicted SMPL-X parameters and groud-truth SMPL-X fits, L_{joint} computes the L_2 loss between the predicted joint coordinates and the ground-truth joint coordinates, L_{bbox} computes the L_2 loss between the predicted and ground-truth bounding box center and size of the face and hands.

Method	MPJPE ↓	MVE ↓	NMJE ↓	NMVE ↓
HMR [22]	180.5	173.6	226.0	217.0
SPIN [25]	175.1	168.7	223.1	216.3
PyMAF [56]	200.2	207.4	168.2	174.2
EFT [21]	196.3	203.6	159.0	165.4
ROMP [45]	116.6	113.8	134.0	130.8
BEV [46]	105.3	100.7	113.2	108.3
Hand4Whole [34]	89.8	91.1	103.2	104.6
MEPO (Ours)	**85.5**	**89.8**	**97.4**	**102.7**

Table 2: Quantitative comparisons with SOTA body-only based methods. We compared our model with existing body-only methods (either single-person or multi-person) to illustrate the effectiveness of our method.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Datasets: We train our model on the Human3.6M [19], MSCOCO [32], MPII [1], and AGORA [37] training dataset, and evaluate our model on the AGORA validation dataset, 3DPW [50], and EHF [39]. AGORA is a synthetic dataset for 3D human pose estimation. The dataset contains 14,000 high-resolution (4K) images for training and 3,000 test images, each containing 5 to 15 people. 3DPW [50] is the first multiple-people dataset that captures human poses and motions in the wild, including 51,000 frames of videos. AGORA and 3DPW are considered challenging datasets because they contain person-person occlusions, environmental occlusions, and camera frame occlusions. EHF [39] dataset is a single-person whole-body dataset including 100 images with SMPL-X annotations and uses a vertex-to-vertex error.

Evaluation Metrics. The evaluation metrics on the AGORA [37] dataset follow the AGORA official benchmark, including Normalized Mean Vertex Error (**NMVE**), Normalized Mean Joint Error (**NMJE**), Mean Vertex Error (**MVE**), and Mean Per Joint Position Error (**MPJPE**) for whole-body, body, right hand, left hand, and face. For 3DPW [50] evaluation, we report MPJPE, Mean Per Vertex Position Error (**MPVPE**), and Procrustes Analysis MPJPE (**PA-MPJPE**). For EHF [39] evaluation, we report PA-MPJPE and MPVPE for whole-body, hands, and face.

4.2 Implementation Details

We implement MEPO with Pytorch [36]. We use OpenPose [6] for instance detection and joint heatmap construction. We use

Method	MPJPE ↓	PA-MPJPE ↓	MPVPE ↓
HMR [22]	130.0	76.7	-
SPIN [25]	121.2	69.9	144.1
ROMP [45]	91.3	54.9	108.3
PIXIE [13]	91.0	61.3	-
Hand4Whole [34]	86.6	54.4	-
MEPO (Ours)	**79.1**	**52.4**	**96.8**

Table 3: Quantitative comparisons with SOTA methods on 3DPW dataset. We compared our model with existing methods on the 3DPW dataset to illustrate the effectiveness of our method in real-world scenarios.

Resnet-50 [18] to process the image crops. Our model is trained and fine-tuned on the Human3.6M [19], MSCOCO [32], MPII [1] and AGORA [37] dataset, and is evaluated on AGORA, 3DPW [50], and EHF [39] validation dataset. The ground-truth depth order levels are obtained from the ground-truth masks as described in Section 3.3. For single-person methods as the baseline, we use OpenPose to detect instances in the image and then pass them to the pose estimation methods for a fair comparison.

4.3 Comparison with State-of-the-Art Methods

AGORA Evaluation. We compare our method against the state-of-the-art methods under the same AGORA evaluation protocol. For fair comparisons, our model uses OpenPose [6] to detect people in the image, which is the same instance detection settings of the image pre-processing for state-of-the-art methods reported in AGORA. The results are given in Table 1. From the table, we can observe that our proposed model, MEPO, performs better than state-of-the-art whole-body pose estimation methods.

As illustrated in Section 3, our model is aware of person-person occlusions in the images, and AGORA includes many occlusions to simulate real-world scenes. Other models that are built for single-person detection are not capable of dealing with complex occlusions.

We further compare our method against the leading edge, body-only pose estimation methods, adhering to the AGORA evaluation protocol. The objective of this comparison is to substantiate that our model surpasses the currently available multi-person pose estimation methods. Given that these existing multi-person methodologies generate body-only models, we compare our model under this body-only setting. Specifically, we use the BodyBranch in our model to predict the body-only 3D human model for fair comparisons. Under body-only settings, we also use the AGORA benchmark to evaluate the models, which calculate NMVE, NMJE, MVE, and MPJPE on body pose predictions. For the performance of baseline models, we use the performance reported by [37]. As shown in Table 2, the proposed MEPO outperforms other baseline methods, i.e., HMR [22], SPIN [25], PyMAF [56], EFT [21], ROMP [45], BEV [46], and Hand4Whole (body-only) [34]. Since AGORA contains complex occlusions, methods that are unaware of occlusions, such as HMR, SPIN, and PyMAF, do not have good overall performance. On the contrary, methods that consider the occlusion issue or depth order issue, such as BEV [46] and our model, have comparatively good performance. These experiment results validate that our approaches are efficient in complex occluded cases and outperform existing multi-person methods.

Methods	PA-MPVPE ↓			MPVPE ↓		
	All	Hands	Face	All	Hands	Face
ExPose [10]	54.5	12.8	5.8	77.1	51.6	35.0
Frankmocap [43]	57.5	12.6	-	107.6	42.8	-
PIXIE [13]	55.0	11.1	**4.6**	89.2	42.8	32.7
Hand4Whole [34]	50.3	10.8	5.8	76.8	39.8	26.1
MEPO (Ours)	**48.4**	**10.7**	5.6	**74.7**	**38.3**	**26.1**

Table 4: Quantitative comparisons with the SOTA whole-body methods on EHF dataset.

Method	MPJPE ↓	MVE ↓	NMJE ↓	NMVE ↓
Baseline	89.1	93.1	103.2	107.0
+ heatmap	86.8	91.9	100.5	106.7
MEPO w/o L_{depth}	**86.4**	**91.1**	**99.4**	**104.8**

Table 5: Ablation results for HENet on AGORA.

3DPW Evaluation. We compare our results with existing pose estimation methods on the 3DPW [50] dataset, which captures human motions in the wild. For a fair comparison with body-only methods, we use the BodyBranch of our model to compare with other methods in the 3DPW dataset. For single-person methods, we use OpenPose to extract instance crops for a fair comparison. The results are shown in Table 3. We compared our model with HMR [22], SPIN [25], ROMP [45], PIXIE [13] and Hand4Whole [34]. Our model outperforms the aforementioned methods.

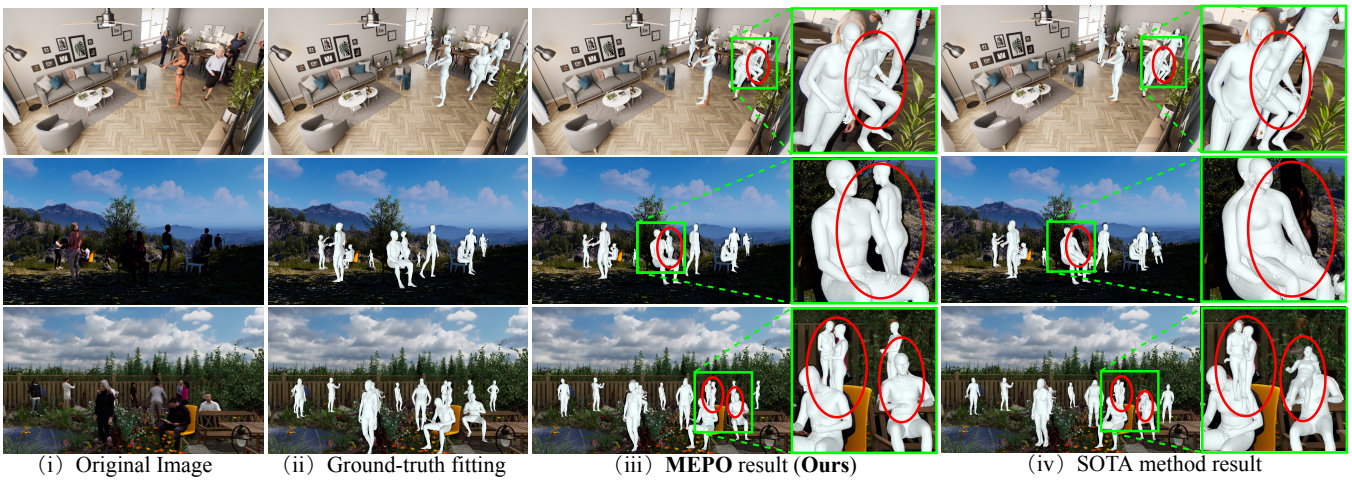
EHF Evaluation. Since our model is a whole-body pose estimation method, we also compare our method with the existing whole-body method on a single-person whole-body pose dataset, EHF. The experiment result is shown in Table 4. We compared our method with the state-of-the-art methods, and our model outperforms these approaches. This illustrates that the design of our proposed method also benefits single-person pose estimation.

We also present the qualitative result of MEPO in Figure 6. The images are from the AGORA [37] and 3DPW [50] dataset. Images contain ambiguous bounding boxes and various occlusions, including person-person occlusions, environmental occlusions, and camera frame occlusions, which is challenging for single-person pose estimation methods. We compare our qualitative results with the leading whole-body method, i.e., Hand4Whole [34]. As circled in Figure 6, it is challenging for the SOTA method to recover human models under such occlusions. However, our model is able to handle such complex situations and provide accurate whole-body representation. The qualitative result shows that our model is robust with crowds and can handle complex person-person occlusions.

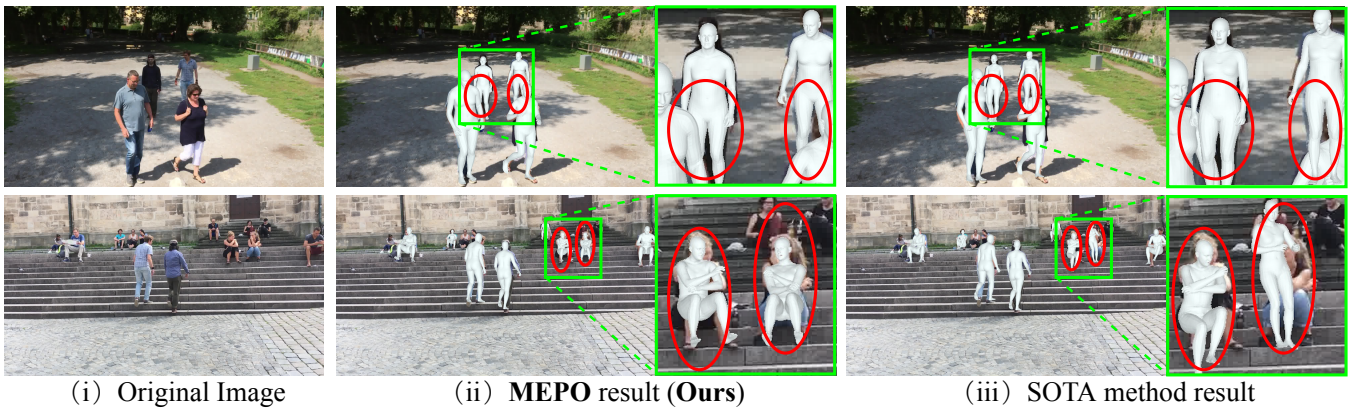
4.4 Ablation Study

To illustrate the essential performance improvement of our methods and the different effects of our methods under different settings, we compare the performance on the AGORA dataset [37] under different experiment settings.

HENet. We first evaluate the effect of the Heatmap Enhancement Net (HENet) on performance. We proceed with a comparative analysis across three implementations: (1) the first implementation is the baseline model without any incorporation of heatmap information; (2) the second implementation directly concatenates the image feature with the heatmap feature; (3) the third implementation integrates the HENet design of our model. The primary objectives



(a) Qualitative results of our method and SOTA method [34] comparisons on AGORA [37] dataset.



(b) Qualitative results of our method and SOTA method [34] comparisons on 3DPW [50] dataset.

Figure 6: Qualitative results of our method and SOTA method [34] comparisons on AGORA [37] and 3DPW [50] dataset. Compared to the SOTA method issues (red circled), our proposed model is robust to person-person occlusions. These results validate that our model performs better under crowded situations.

Method	MPJPE ↓	MVE ↓	NMJE ↓	NMVE ↓
MEPO w/o L_{depth}	86.4	91.1	99.4	104.8
MEPO	**85.8**	**89.1**	**97.1**	**102.1**

Table 6: Ablation results for depth order loss.

of this experimental setup are twofold. First, we aim to demonstrate the critical role of heatmap information in our methodology. Second, we intend to establish that our proposed HENet enhances performance beyond what is achieved by merely concatenating image and heatmap features. We evaluate these models via the AGORA evaluation benchmark.

The experiment results are shown in Table 5. We can observe that in both experiments, the two strategies using heatmap both have better performance, while our proposed HENet structure performs better than simply concatenating the image and heatmap features together. This shows that heatmap is beneficial for human pose estimation. Furthermore, it proves that our proposed HENet efficiently utilizes the information delivered by the joint heatmap.

Depth Order Loss. To investigate whether the depth order loss leads to positive effects on the model performance, we also designed a comparative experiment. We compared our model before and after adding the depth order loss. The results are listed in Table 6. As shown in Table 6, after applying losses, our model has a lower MPJPE, MVE, NMJE, and NMVE than the baseline.

5 CONCLUSION AND FUTURE WORK

In this paper, we present MEPO, a novel monocular, multi-person, 3D whole-body human pose recovery method to address the problem of expressive pose estimation of multiple people. We propose to use HENet to extract heatmap information efficiently. The model is depth-aware by introducing depthmap and depth order loss. Our method has proved efficient while dealing with either crowd scenes or single-person cases. Our method outperforms existing whole-body pose estimation methods and multi-person pose estimation methods. In the future, we plan to explore the possibilities of the one-stage method for multi-person whole-body pose estimation.

Acknowledgments: This research is partially supported by NSF IIS-2309073 and ECCS-2123521. This article solely reflects the opinions and conclusions of its authors and not the funding agencies.

REFERENCES

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2014. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition*. 3686–3693.

[2] Dragomir Anguelov, Praveen Srinivasan, Daphne Koller, Sebastian Thrun, Jim Rodgers, and James Davis. 2005. Scape: shape completion and animation of people. In *ACM SIGGRAPH 2005 Papers*. 408–416.

[3] Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

[4] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. 2016. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *European conference on computer vision*. Springer, 561–578.

[5] Grigore C Burdea and Philippe Coiffet. 2003. *Virtual reality technology*. John Wiley & Sons.

[6] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7291–7299.

[7] Julie Carmigniani, Borko Furht, Marco Anisetti, Paolo Ceravolo, Ernesto Damiani, and Misa Ivkovic. 2011. Augmented reality technologies, systems and applications. *Multimedia tools and applications* 51, 1 (2011), 341–377.

[8] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. 2016. Single-image depth perception in the wild. *Advances in neural information processing systems* 29 (2016).

[9] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. 2021. 3DCrowdNet: 2D Human Pose-Guided3D Crowd Human Pose and Shape Estimation in the Wild. *arXiv preprint arXiv:2104.07300* (2021).

[10] Vasileios Choutas, Georgios Pavlakos, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2020. Monocular Expressive Body Regression through Body-Driven Attention. In *European Conference on Computer Vision (ECCV)*. 20–40. https://expose.is.tue.mpg.de

[11] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 2019. Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[12] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 2020. 3d morphable face models—past, present, and future. *ACM Transactions on Graphics (TOG)* 39, 5 (2020), 1–38.

[13] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael Black. 2021. Collaborative Regression of Expressive Bodies using Moderation. In *International Conference on 3D Vision (3DV)*. 792–804. https://doi.org/10.1109/3DV53792.2021.00088

[14] Yao Feng, Vasileios Choutas, Timo Bolkart, Dimitrios Tzionas, and Michael J Black. 2021. Collaborative Regression of Expressive Bodies using Moderation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 792–804.

[15] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. 2021. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (TOG)* 40, 4 (2021), 1–13.

[16] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. 2020. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3196–3206.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

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

[19] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 2013. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence* 36, 7 (2013), 1325–1339.

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

[21] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. 2021. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *2021 International Conference on 3D Vision (3DV)*. IEEE, 42–52.

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

[23] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. 2022. Occluded human mesh recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1715–1725.

[24] Seung-Wook Kim, Hyong-Keun Kook, Jee-Young Sun, Mun-Cheon Kang, and Sung-Jea Ko. 2018. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 234–250.

[25] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. 2019. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2252–2261.

[26] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael Bronstein, and Stefanos Zafeiriou. 2019. Single image 3d hand reconstruction with mesh convolutions. *arXiv preprint arXiv:1905.01326* (2019).

[27] Timothy E Lee, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Oliver Kroemer, Dieter Fox, and Stan Birchfield. 2020. Camera-to-robot pose estimation from a single image. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 9426–9432.

[28] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3383–3393.

[29] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. 2022. Interacting Attention Graph for Single Image Two-Hand Reconstruction. *arXiv preprint arXiv:2203.09364* (2022).

[30] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.* 36, 6 (2017), 194–1.

[31] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

[32] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)* 34, 6 (2015), 1–16.

[34] Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. 2022. Accurate 3D Hand Pose Estimation for Whole-Body 3D Human Mesh Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2308–2317.

[35] Gyeongsik Moon and Kyoung Mu Lee. 2020. Pose2pose: 3d positional pose-guided 3d rotational pose prediction for expressive 3d human pose and mesh estimation. *arXiv preprint arXiv:2011.11534* (2020).

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[37] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. 2021. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

[38] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10975–10985.

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

[40] Lionel Pigou, Sander Dieleman, Pieter-Jan Kindermans, and Benjamin Schrauwen. 2014. Sign language recognition using convolutional neural networks. In *European Conference on Computer Vision*. Springer, 572–578.

[41] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017).

[42] Javier Romero, Dimitrios Tzionas, and Michael J Black. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610* (2022).

[43] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. 2020. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324* (2020).

[44] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. 2023. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1972–1981.

[45] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Michael J Black, and Tao Mei. 2021. Monocular, one-stage, regression of multiple 3d people. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision. 11179–11188.

[46] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13243–13252.

[47] Bugra Tekin, Federica Bogo, and Marc Pollefeys. 2019. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4511–4520.

[48] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. 2018. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2549–2559.

[49] Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1126–1135.

[50] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.

[51] Bo Wan, Desen Zhou, Yongfei Liu, Rongjie Li, and Xuming He. 2019. Pose-aware multi-level feature network for human object interaction detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 9469–9478.

[52] Haoran Wang, Yue Fan, Zexin Wang, Licheng Jiao, and Bernt Schiele. 2018. Parameter-free spatial attention network for person re-identification. *arXiv preprint arXiv:1811.12150* (2018).

[53] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. 2020. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. In *European Conference on Computer Vision*. Springer, 380–397.

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19.

[55] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. 2021. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5746–5756.

[56] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. 2021. Pymaf: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11446–11456.

[57] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. 2021. Body meshes as points. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 546–556.

[58] Yang Zheng, Ruizhi Shao, Yuxiang Zhang, Tao Yu, Zerong Zheng, Qionghai Dai, and Yebin Liu. 2021. Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6239–6249.

[59] Yuxiao Zhou, Marc Habermann, Ikhsanul Habibie, Ayush Tewari, Christian Theobalt, and Feng Xu. 2021. Monocular real-time full body capture with inter-part correlations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4811–4822.