Instance-aware Contrastive Learning for Occluded Human Mesh Reconstruction

Mi-Gyeong Gwon[1] Gi-Mun Um[2] Won-Sik Cheong[2] Wonjun Kim[1*]
[1]Konkuk University [2]Electronics and Telecommunications Research Institute
{kmk3942,wonjkim}@konkuk.ac.kr {gmum,wscheong}@etri.re.kr

Abstract

A simple yet effective method for occlusion-robust 3D human mesh reconstruction from a single image is presented in this paper. Although many recent studies have shown the remarkable improvement in human mesh reconstruction, it is still difficult to generate accurate meshes when person-to-person occlusion occurs due to the ambiguity of who a body part belongs to. To address this problem, we propose an instance-aware contrastive learning scheme. Specifically, joint features belonging to the target human are trained to be proximate with the center feature (i.e., feature extracted from the body center position). On the other hand, center features of different human instances are forced to be far apart so that joint features of each person can be clearly distinguished from others. By interpreting the joint possession based on such contrastive learning scheme, the proposed method easily understands the spatial occupancy of body parts for each person in a given image, thus can reconstruct reliable human meshes even with severely overlapped cases between multiple persons. Experimental results on benchmark datasets demonstrate the robustness of the proposed method compared to previous approaches under person-to-person occlusions. The code and model are publicly available at: https://github.com/DCVL-3D/InstanceHMR_release.

1. Introduction

Recently, 3D human mesh reconstruction has become a major research topic with the increasing demand in entertainment areas such as AR/VR and sports broadcasting. In line with this trend, considerable efforts have been made to predict reliable meshes from a single image by using deep learning techniques. Specifically, model-based methods, which aim to reconstruct human poses and shapes by estimating parameters of the skinned multi-person linear (SMPL) model [32], have been introduced. Since the problem of inferring a large number of vertices is simplified to the problem of estimating a few parameters through various network architectures, model-based approaches have been actively explored while bringing the significant progress in human mesh reconstruction [2, 4, 15, 18, 19, 21, 24, 25, 37, 46]. On the other hand, model-free approaches also have drawn a lot of attentions due to their ability to express local details of mesh surfaces by directly regressing every vertex without any constraint of the parameter space. In particular, the transformer architecture has been widely adopted to reconstruct accurate meshes by considering the relationship between vertices in a global manner [5, 17, 27, 28, 44, 45, 50, 51].

Despite meaningful advances in the field of 3D human mesh reconstruction, person-to-person occlusions frequently occurring in crowd scenes still remain a challenging issue due to ambiguities of possession for body parts. To resolve this issue, several approaches have recently begun to be introduced. For example, there have been attempts to estimate discriminative locations of each person in the 2D space and utilize them to extract features for individual mesh regression [7, 16, 38]. Reasoning the depth order of multiple people also has been considered to understand the relationship between overlapped subjects in a given image [11, 39, 47]. Some methods first predict the occlusion-robust 3D skeletons, which form reliable poses even for invisible body parts, and then lift them to mesh structures [31]. Even though considerable efforts have been consistently devoted to this task, attempts to exploit contextual cues of the occluded area, which is occupied by each person, are still insufficient.

In this paper, a new perspective for effectively handling person-to-person occlusions is presented. In the scene where people are overlapped, one of the biggest obstacles is the uncertainty about the ownership of a body part. While it is relatively easy to semantically understand which area in a given image corresponds to which type of body part (e.g., hand or foot), it can be challenging to determine who owns that body part. This makes the model confused and eventually yields unnatural reconstruction results. To deal with this problem, we propose to guide the network to learn discriminative representations of body parts for each person by our instance-aware contrastive learning scheme. Specifically, two novel feature maps, i.e., center-aligned instance

*corresponding author

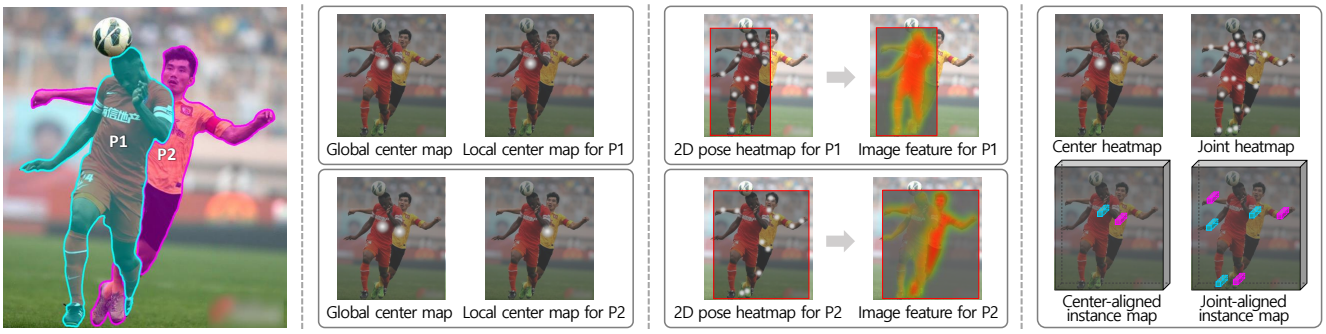


Figure 1. (a) An example of person-to-person occlusion. (b), (c), and (d) Visualizations of the information adopted to guide the feature encoding for human mesh reconstruction in OCHMR [16], 3DCrowdNet [7], and the proposed method, respectively. For better visibility, all joints are represented in a single image at the 2D pose heatmap of (c) and the joint heatmap of (d), and only a few of joint features (e.g., right hand, left elbow, and right foot) are shown in the joint-aligned instance map of (d).

map and joint-aligned instance map, are encoded in the proposed architecture, which can describe the possession relationship between all human instances and their visible body joints. These maps learn to represent the identity information of each human instance by forcing the center feature to pull joint features of the same person while pushing out center features of non-target persons in the latent space. Note that center features and joint features are sampled at the body center position and joint locations in the center-aligned instance map and the joint-aligned instance map, respectively, as shown in Fig. 1(d). Such identity representations make it easy to distinguish which instance each joint belongs to, thus the model can better understand the occluded context and determine the area of the image to focus on for occluded human mesh reconstruction. A comparison between previous approaches and the proposed method for handling person-to-person occlusions is illustrated in Fig. 1. In [16], relative positions of multiple persons are given to the model in the form of global and local center maps as shown in Fig. 1(b). However, since the center position is an extremely small amount of information compared to the complex body structure, the model may hardly figure out details of body part regions in occlusion situations. On the other hand, the spatial guidance of the body region for the target person is provided to the network based on the 2D pose heatmap in [7]. However, features of non-target person areas also can be unnecessarily fused into features for target mesh reconstruction (see the bottom part of Fig. 1(c)). Unlike such previous approaches, the proposed method clearly indicates the possession of all body parts by expressing the identity information at joint locations as well as body center positions via the instance-aware contrastive learning scheme as shown in Fig. 1(d). Moreover, meshes for every person can be reliably reconstructed in a single stream (i.e., without using the bounding box) by considering the mutual relationship among the spatial occupancy of each person's body part. The main contributions of this paper are summarized as follows:

- We propose to indicate the spatial occupancy of body parts corresponding to each human instance for occlusion-robust 3D human mesh reconstruction. With the awareness of the body part possession of every person, the model can easily understand the person-to-person occlusion and successfully generate human meshes with natural poses.
- We represent the relationship between human instances and their spatial keypoints (i.e., body centers and joints) at the feature level via the proposed instance-aware contrastive learning scheme. Specifically, by guiding the center feature of the target person to push out that of non-target persons and pull corresponding joint features in the latent space, identities of different persons can be distinctively embedded into center and joint features.

2. Related Works

In this Section, we present a systematic review of previous methods for monocular 3D human mesh reconstruction and explore various approaches for handling person-to-person occlusions.

2.1. 3D Human Mesh Reconstruction

Deep learning-based methods for 3D human mesh reconstruction can be divided into two main groups, i.e., model-based and model-free methods. The former aims to estimate pose and shape parameters of the SMPL [2] model. In the early stage, Kanazawa *et al*. [15] proposed to regress SMPL parameters by using the end-to-end convolutional neural network and devised the adversarial loss for plausible mesh generation. Kolotouros *et al*. [19] designed the optimization loop that can be combined with end-to-end regression frameworks such as [15], to further refine reconstruction results. Moreover, Zhang *et al*. [46] tried to rectify meshes by iteratively feeding pyramidal mesh-aligned features into the mesh regression network. A bottom-up framework, which predicts meshes of multiple persons in a given image dur-

ing a single inference, was introduced by Sun *et al*. [38] for the first time. Specifically, they proposed to extract SMPL parameters by the sampling process based on body center positions. Li *et al*. [26] incorporated the bounding box location into the network for accurately estimating the global orientation. To reconstruct the human mesh that is aligned well to the image pixel, Shetty *et al*. [37] proposed to first predict vertices on the input image plane, and then compute SMPL parameters by comparing them with the template mesh part by part based on inverse kinematics. Cho *et al*. [4] attempted to learn the consistency of human pose and shape in arbitrary view directions by using neural feature fields. In contrast to such model-based approaches, model-free methods directly estimate 3D coordinates of mesh vertices. As a pioneer, Kolotouros *et al*. [20] adopted the graph convolutional neural network (GraphCNN) to effectively transform image features into vertex coordinates. Choi *et al*. [6] also utilized GraphCNN to lift estimated 2D and 3D poses to the 3D mesh. Meanwhile, Lin *et al*. [27] presented a new direction in model-free approaches, by employing the transformer encoder with the dimension reduction technique for mesh regression. They further improved the performance by incorporating GraphCNN into the transformer encoder block [28]. To alleviate the computational burden of the transformer-based architecture, Cho *et al*. [5] proposed to disentangle image features and geometric parameters of the human body based on the cross-attention module. Kim *et al*. [17] designed a point-guided feature sampling scheme to effectively extract vertex-relevant features for the transformer-based mesh regressor. More recently, Foo *et al*. [8] adopted the diffusion model for the human mesh reconstruction by leveraging the pose heatmap in the diffusion process to enable the mesh estimation conditioned on the input image.

2.2. Handling Person-to-Person Occlusions

Even though many studies mentioned above have brought significant advances for monocular 3D human mesh reconstruction, they still suffer from ambiguities driven by person-to-person occlusions, which often occur in real-world environments. To cope with this problem, there have been various technical attempts. For example, Jiang *et al*. [11] introduced a new loss function that minimizes the discrepancy between the human segmentation masks and projected meshes to arrange the depth order of the people. In a similar vein, Zhang *et al*. [47] tried to learn the ordinal relation by supervising the difference between depth values of multiple persons. Instead of simply guiding the network via loss functions, Sun *et al*. [39] incorporated the bird's-eye-view body center heatmap into the pipeline of mesh estimation to consider the depth cue. Yang *et al*. [43] adopted to synthesize the other people to the input data for consideration of more variants from person-to-person occlusions.

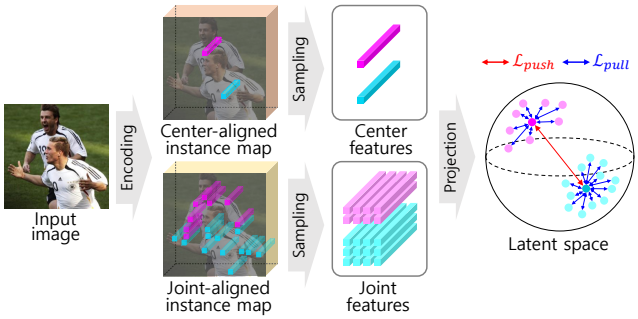


Figure 2. The detailed process of the instance-aware contrastive learning scheme. Note that the ground truth of body center and joint positions is used for the feature sampling.

To specify the position of the target person, Khirodkar *et al*. [16] introduced the context normalization block, which fuses global and local center heatmaps with corresponding features. Choi *et al*. [7] utilized the 2D pose heatmap to focus on the feature belonging to the area of the target person in crowded scenes. Cha *et al*. [3] attempted to refine multiple human meshes based on inter-person relations that are computed by the transformer architecture. Liu *et al*. [31] proposed to apply the knowledge transfer technique to the 3D keypoint detection [49] for successfully reasoning invisible body parts under occlusions. Li *et al*. [22] proposed a body center attention mechanism to consider spatial-temporal relations of multiple persons at the pixel level. To complement the lack of the training data under multi-person scenes, Sun *et al*. [40] tried to synthesize plausible crowded examples by putting human samples at appropriate places of the given image with a reasonable scale according to the scene context. Moreover, they also proposed to learn the consistency between features extracted from the same human sample with and without occlusions.

3. Proposed Method

The proposed method aims to learn the spatial occupancy of body parts for each person by using the instance-aware contrastive learning scheme as shown in Fig. 2. By incorporating the identity information into encoded features, the proposed method can achieve the robust performance against person-to-person occlusions. The overall architecture of the proposed method is illustrated in Fig. 3.

3.1. Instance-aware Contrastive Learning

In the bottom-up approach for human mesh reconstruction, it is important to grasp the individual cue between multiple persons. To this end, we guide the process of mesh estimation based on the regional distinction of each human area, i.e., 24 body joints defined by the SMPL model [2]. Specifically, the backbone feature is encoded into two different feature maps, i.e., center-aligned instance map and joint-aligned instance map, through the corresponding network

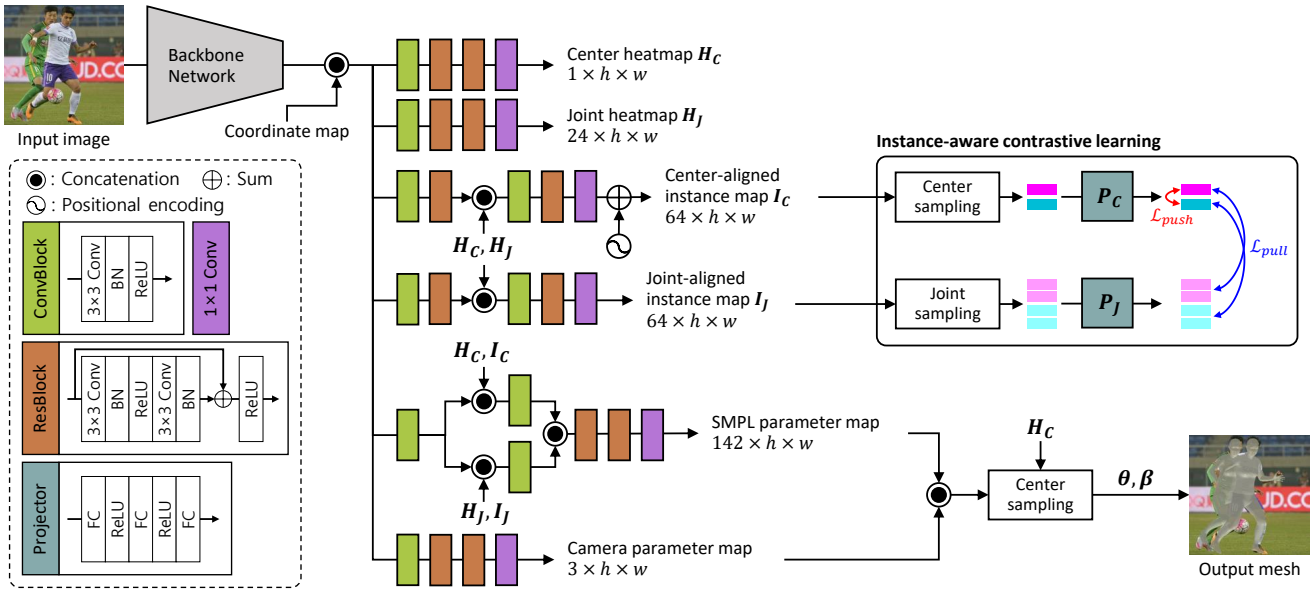


Figure 3. The overall architecture of the proposed method for occluded human mesh reconstruction. h and w are set to 64. θ and β denote pose and shape parameters of the SMPL model. Note that instance-aware contrastive learning is conducted only in the training phase.

branch as shown in Fig. 3. Note that the predicted center heatmap H_C and joint heatmap H_J are adopted in this encoding process to provide the spatial information of body centers and joints. In addition, the positional encoding is also applied to the center-aligned instance map to prevent ambiguous representations between different people who have similar appearances. After that, center features and joint features are sampled at body center locations and visible joint positions of the center-aligned instance map and the joint-aligned instance map, respectively. These sampled features are subsequently projected into the latent space by using the nonlinear projector, which consists of fully connected layers and ReLUs (see P_C and P_J in Fig. 3).

Now, such projected features are fed into our instance-aware contrastive learning scheme. The detailed process is shown in Fig. 2. First, the center feature of the target person (i.e., anchor) is forced to push out those of other persons. By placing different human instances far apart from each other in the latent space, center features can implicitly represent the identity information for each person in a discriminative way. Second, the center feature pulls joint features belonging to the same person to a close distance. This facilitates the joint feature to express the human instance to which it belongs, by embedding the target identity information. To this end, two loss functions are designed based on the cosine similarity, which is suitable for measuring the directional coherence between embedding vectors, as follows:

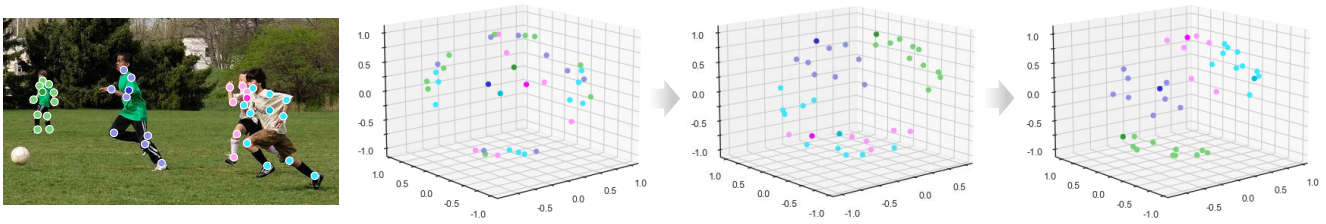
$$\begin{aligned} \mathcal{L}_{push}(a,b) &= 1 + \frac{a \cdot b}{\|a\|\|b\|}, \\ \mathcal{L}_{pull}(a,b) &= 1 - \frac{a \cdot b}{\|a\|\|b\|}. \end{aligned} \tag{1}$$

Specifically, $\mathcal{L}_{push}$ is used to maximize the distance between center features of different people and $\mathcal{L}_{pull}$ is used to minimize the distance between the center feature and the joint feature of the same person in the latent space. By considering every possible center-to-center and center-to-joint combination, the proposed contrastive loss is defined as follows:

$$\begin{aligned} \mathcal{L}_{cont} = &\frac{2}{N(N-1)} \sum_{i \neq j}^{N} \mathcal{L}_{push}(u_i, u_j) \\ &+ \frac{1}{N} \sum_{i=1}^{N} \frac{1}{K_i} \sum_{k=1}^{K_i} \mathcal{L}_{pull}(u_i, v_i^k), \end{aligned} \tag{2}$$

where u and v indicate the center feature and the joint feature, respectively. N and K_i denote the number of people and the number of visible joints belonging to the i-th person in the input image, respectively. The effect of the proposed instance-aware contrastive learning scheme is shown in Fig. 4. As can be seen, the identity information is embedded into latent features during our instance-aware contrastive learning. Consequently, it is thought that both center-aligned instance map and joint-aligned instance map play an important role to describe the possession relationship between human instances and their visible joints. Note that the instance-aware contrastive learning part is conducted only in the training phase.

In order to utilize this possession cue in the process of mesh regression, we incorporate the center-aligned instance map and the joint-aligned instance map into the network branch for estimating the SMPL parameter map (see the second branch from bottom in Fig. 3). To accurately indicate valid locations, the center heatmap and the joint



(a) (b) (c) (d)

Figure 4. A visualization example of center features and joint features in the latent space. The feature dimension is reduced from 64 to 3 by using [33] for the visualization. (a) Input image with marks for positions of body centers (dark) and joints (bright). Note that points belonging to the same person are represented as the same color attribute. The effect of the proposed instance-aware contrastive learning scheme is shown sequentially, i.e., (b) before training, (c) middle of training, and (d) after training.

heatmap are also employed for this concatenation process. Finally, pose and shape parameters of each person are sampled from the SMPL parameter map according to the predicted center position in a similar way of [38]. It is noteworthy that the model can precisely consider the occupancy of each human instance via such identity representation in the feature space, which leads to generation of reliable human meshes even with various person-to-person occlusions. An example of occluded human mesh reconstruction is shown in the bottom right of Fig. 3.

3.2. Loss functions

The proposed method is trained based on nine loss terms, i.e., contrastive loss $\mathcal{L}_{cont}$ (which is explained in subsection 3.1), body center heatmap loss $\mathcal{L}_{center}$, joint heatmap loss $\mathcal{L}_{joint}$, pose loss $\mathcal{L}_{pose}$, shape loss $\mathcal{L}_{shape}$, Mixture Gaussian prior loss $\mathcal{L}_{prior}$, 3D joint loss $\mathcal{L}_{3d}$, Procrustes-aligned 3D joint loss $\mathcal{L}_{pa3d}$, and 2D joint loss $\mathcal{L}_{2d}$. First, the focal loss [30, 38] is adopted to train the body center heatmap H_C and the joint heatmap H_J, which is defined as follows:

$$\begin{aligned}
\mathcal{L}_{focal}(H,\tilde{H},\tilde{H}^{'}) &= -\frac{\mathcal{L}_{pos}(H,\tilde{H}^{'})+\mathcal{L}_{neg}(H,\tilde{H},\tilde{H}^{'})}{\sum \tilde{H}^{'}},\\
\mathcal{L}_{neg}(H,\tilde{H},\tilde{H}^{'}) &= \log(1-H)H^2(1-\tilde{H})^4(1-\tilde{H}^{'}),\\
\mathcal{L}_{pos}(H,\tilde{H}^{'}) &= \log(H)(1-H)^2\tilde{H}^{'},
\end{aligned} \tag{3}$$

where H and $\tilde{H}$ denote a single heatmap and the corresponding ground truth, respectively. $\tilde{H}^{'}$ indicates the binary map that is marked on the location of the positive class in $\tilde{H}$. By using the focal loss, $\mathcal{L}_{center}$ and $\mathcal{L}_{joint}$ are formulated as follows:

$$\begin{aligned}
\mathcal{L}_{center} &= \mathcal{L}_{focal}(H_C,\tilde{H}_C,\tilde{H}_C^{'}),\\
\mathcal{L}_{joint} &= \frac{1}{\mathcal{J}}\sum_{j=1}^{\mathcal{J}}\mathcal{L}_{focal}(H_J^j,\tilde{H}_J^j,\tilde{H}_J^{'j}),
\end{aligned} \tag{4}$$

where $\mathcal{J}$ is the number of joint heatmaps which is set to 24 according to the SMPL model. The remaining loss terms are used to directly supervise the process of human mesh reconstruction. In particular, $\mathcal{L}_{pose}$ and $\mathcal{L}_{shape}$ compute the difference between predicted pose and shape parameters and the corresponding ground truth based on L_2 loss, respectively [15, 20]. Mixture Gaussian prior loss $\mathcal{L}_{prior}$ is employed to attain the naturalness of body meshes as introduced in [2]. In our work, Euclidean distance is utilized to calculate the spatial gap in the coordinate system for following three loss terms. The distance between estimated 3D joints and the corresponding ground truth is computed by $\mathcal{L}_{3d}$ [15, 20]. $\mathcal{L}_{pa3d}$ indicates $\mathcal{L}_{3d}$ after Procrustes alignment is conducted to the reconstructed mesh [38]. The coordinates of joints projected onto the 2D space are used to measure $\mathcal{L}_{2d}$ [15, 20]. Finally, the total loss function is defined by using the weighted sum of all the aforementioned loss terms as follows:

$$\begin{aligned}
\mathcal{L}_{total} = {} & \lambda_{cont}\mathcal{L}_{cont}+\lambda_{center}\mathcal{L}_{center}+\lambda_{joint}\mathcal{L}_{joint}\\
& +\lambda_{pose}\mathcal{L}_{pose}+\lambda_{shape}\mathcal{L}_{shape}+\lambda_{prior}\mathcal{L}_{prior}\\
& +\lambda_{3d}\mathcal{L}_{3d}+\lambda_{pa3d}\mathcal{L}_{pa3d}+\lambda_{2d}\mathcal{L}_{2d},
\end{aligned} \tag{5}$$

where λ_{cont}, λ_{center}, λ_{joint}, λ_{pose}, λ_{shape}, λ_{prior}, λ_{3d}, λ_{pa3d}, and λ_{2d} are the balancing factor for each loss term, which are set to 50, 160, 50, 80, 6, 1.6, 200, 360, and 400, respectively.

4. Experimental Results

4.1. Implementation Details

The proposed method is implemented on the PyTorch framework with an Intel E5-1650@3.60GHz CPU and two NVIDIA GeForce RTX 3090 GPUs. All the parameters of the proposed network are updated by the Adam optimizer, where momentum factors are set to 0.9 and 0.999, respectively. We use the batch size of 64 and the learning rate is set to 5×10^{-6} during 30 training epochs. Input images are randomly cropped and rotated before they are resized to the resolution of 512×512 pixels by using zero padding.

4.2. Benchmark Datasets

The proposed method is trained by using three 3D human pose datasets (i.e., Human3.6M [10], MPI-INF-3DHP [34],

Figure 5. Results of 3D human mesh reconstruction on the CrowdPose [23] dataset. From top to bottom: input images, results by ROMP [38], 3DCrowdNet [7], and the proposed method.

Methods	3DPW-PC			OCHuman					CrowdPose		
	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)	AP(↑)	AP^{50}(↑)	AP^{75}(↑)	AP^{M}(↑)	AP^{L}(↑)	AP(↑)	AP^{50}(↑)	AP^{75}(↑)
SPIN* [19]	129.6	82.6	157.6	12.7	46.8	19.4	17.8	26.2	16.4	40.1	10.6
PyMAF* [46]	126.7	81.3	154.3	14.3	48.7	21.5	18.0	28.7	17.4	42.7	13.0
ROMP* [38]	119.7	79.7	152.8	15.6	55.0	23.6	18.7	30.0	18.9	44.6	13.8
ROMP [38]	115.6	75.8	147.5	19.8	56.2	25.0	19.3	32.9	28.5	58.8	24.7
OCHMR* [16]	117.5	77.1	149.6	24.8	60.7	**28.6**	22.3	34.2	21.4	48.3	16.5
CLIFF* [26]	–	–	–	23.2	49.8	20.4	–	–	27.4	51.5	25.9
HMDiff* [8]	114.2	**73.5**	143.1	–	–	–	–	–	–	–	–
CoordFormer* [22]	101.5	79.3	–	–	–	–	–	–	–	–	–
Ours*	**99.9**	76.3	**126.4**	30.5	75.8	17.3	**35.1**	30.5	28.5	61.6	23.2
Ours	102.0	77.2	131.0	**34.9**	**80.0**	24.0	27.1	**35.0**	**31.6**	**65.7**	**26.7**

Table 1. Performance comparison of occluded human mesh reconstruction based on 3DPW-PC, OCHuman, and CrowdPose datasets (best results and second results are shown in bold and underlined, respectively). Note that * denotes the performance without using the CrowdPose dataset for training.

and MuCo-3DHP [35]) and four 2D human pose datasets (i.e., MPII [1], LSP [12], COCO [29], and CrowdPose [23]). In particular, only 2D pose labels are used for [23] while pseudo mesh labels, which are generated by [14], are employed to utilize other 2D human pose datasets [1, 12, 29] for training.

For the performance evaluation of the proposed method, four person-to-person occlusion datasets (i.e., 3DPW-PC [41], OCHuman [48], CrowdPose [23], and CMU-Panoptic [13]) are adopted. Specifically, the 3DPW-PC dataset is a subset of the 3DPW dataset where images taken under person-to-person occlusions are collected according to [38]. The OCHuman dataset consists of severe occlusion cases between multiple persons while the CrowdPose dataset contains various person-to-person occlusions occurring in crowded scenes. Moreover, the CMU-Panoptic dataset includes multi-person sequences with diverse inter-person interactions. Based on such occlusion-oriented benchmark datasets, the performance of the proposed method will be evaluated and analyzed in-depth in the following subsection.

4.3. Performance Evaluation

Quantitative Evaluation. To show the effectiveness of the proposed method, we compare ours with previous meth-

Figure 6. More results of 3D human mesh reconstruction by the proposed method on 3DPW-PC [41] (1st row) and OCHuman [48] (2nd row) datasets.

Methods	Haggl.	Mafia	Utim.	Pizza	Mean
SPIN [19]	124.3	132.4	150.4	153.5	133.1
CRMH [11]	129.6	133.5	153.0	156.7	143.2
BMP [47]	120.4	132.7	140.9	147.5	135.4
PARE [18]	143.1	193.1	219.8	190.4	186.6
ROMP [38]	111.8	129.0	148.5	149.1	134.6
OCHMR [16]	115.5	123.7	142.6	150.6	133.1
3DCrowdNet [7]	109.6	135.9	**129.8**	135.6	127.6
Ours	**109.1**	**122.5**	137.8	**135.0**	**126.1**

Table 2. Performance comparison by MPJPE based on the CMU-Panoptic dataset (best results and second results are shown in bold and underlined, respectively).

Methods	MPJPE(↓)	PA-MPJPE(↓)	MPVPE(↓)
METRO [27]	77.1	47.9	88.2
PARE [18]	74.5	46.5	88.6
ROMP [38]	76.7	47.3	93.4
MeshGraphormer [28]	74.7	45.6	87.7
FastMETRO [5]	73.5	44.6	84.1
CLIFF [26]	**72.0**	45.7	85.3
PointHMR [17]	73.9	44.9	85.5
ImpHMR [4]	74.3	45.4	87.1
HMDiff [8]	72.7	44.5	82.4
CoordFormer [22]	79.4	46.5	94.4
Ours (ResNet-50)	78.0	47.8	85.6
Ours (HRNet-32)	73.2	**44.3**	**80.3**

Table 3. Performance comparison based on the 3DPW dataset by following the protocol 3, i.e., fine-tuned on 3DPW (best results and second results are shown in bold and underlined, respectively).

ods for 3D human mesh reconstruction, i.e., SPIN [19], CRMH [11], BMP [47], PyMAF [46], PARE [18], ROMP [38], OCHMR [16], 3DCrowdNet [7], HMDiff [8], and CoordFormer [22], based on person-to-person occlusion benchmarks. To evaluate the proposed method on 3D human mesh datasets, three metrics are adopted, i.e., mean per joint position error (MPJPE) [10], Procrustes-aligned mean per joint position error (PA-MPJPE) [52], and mean per vertex position error (MPVPE) [36]. Specifically, MPJPE indicates the average value of the Euclidean distance between the predicted 3D joint and the corresponding ground truth. PA-MPJPE means MPJPE that is calculated after applying the Procrustes analysis to the estimated body mesh. MPVPE is computed by averaging the Euclidean distance between the predicted vertex and the corresponding ground truth. For 2D human pose datasets, the performance is evaluated by using the average precision (AP), which is calculated based on the object keypoint similarity (OKS) [29]. In addition, AP according to different threshold values (i.e., AP^{50} and AP^{75}) and scales of the human object (i.e., AP^{M} and AP^{L}) is also adopted for the performance comparison. Note that ResNet-50 [9] is employed as the backbone network of the proposed method, which is the default setting for the performance report in this subsection.

First of all, the performance comparison based on 3DPW-PC, OCHuman, and CrowdPose datasets is shown in Table 1. As can be seen, the proposed method shows the meaningful performance improvement compared to previous approaches. Specifically, MPJPE and MPVPE on the 3DPW-PC dataset prove that our model can successfully estimate both pose and shape for occluded human mesh reconstruction. Moreover, the proposed method outperforms all state-of-the-art methods in terms of AP and AP^{50} with the significant performance gain on OCHuman and CrowdPose datasets. In Table 2, MPJPE of the proposed method on the CMU-Panoptic dataset is compared with others. We can see that the proposed method yields the reliable performance in most sequences while achieving the best result in the average MPJPE (see the rightmost column of Table 2). From the result reported in Tables 1 and 2, we could confirm that the proposed instance-aware contrastive learning scheme is effective to reconstruct human meshes under various person-to-person occlusions. Furthermore, we also evaluate the proposed method on the entire test set of 3DPW

Methods	MPJPE(↓)	PA-MPJPE(↓)
Baseline (i.e., w/o center and joint features)	104.2	80.2
Ours†	102.2	78.2
Ours	102.0	77.2

Table 4. Performance analysis according to the change of the network architecture. † denotes the performance without applying positional encoding to the center-aligned instance map.

Feature Dimension	MPJPE(↓)	PA-MPJPE(↓)
32	103.5	78.3
64	102.0	77.2
128	103.1	78.0
256	104.2	78.6

Table 5. Performance variations according to the change of the feature dimension for the center-aligned instance map and the joint-aligned instance map.

by following the protocol 3 [41]. In particular, we conduct the experiment by using two different backbone networks, i.e., ResNet-50 [9] and HRNet-32 [42], and the corresponding results are shown in Table 3. As can be seen, the proposed method performs robustly in a variety of real-world environments, not just for occlusion cases.

Qualitative Evaluation. The qualitative comparison of the proposed method with ROMP [38] and 3DCrowdNet [7] based on the CrowdPose dataset is shown in Fig. 5. As can be seen, the proposed method can successfully reconstruct multiple human meshes in complicated person-to-person occlusions. Specifically, the reconstructed mesh is well algined to the target person without confusion with other persons (see the left two examples in Fig. 5). This is because the proposed method is able to effectively utilize the identity information of each person in the feature space. The fourth and fifth columns of Fig. 5 show the robustness of our method in crowded scenes. Furthermore, the proposed method works reliably even under severe occlusions between multiple persons as shown in the rightmost three examples in Fig. 5. More reconstruction results by the proposed method on 3DPW-PC and OCHuman datasets are also shown in Fig. 6. Based on this, it is thought that our instance-aware contrastive learning scheme is greatly helpful in understanding the occlusion context by overlapped persons and generating the plausible human mesh.

4.4. Ablation Study

In this subsection, several comparative experiments are conducted to demonstrate the effectiveness of the proposed method. The performance for all the experiments in this subsection is evaluated on the 3DPW-PC [41] dataset. First

(a) (b) (c)

Figure 7. (a) Input image. (b) Reconstruction result by baseline. (c) Reconstruction result by the proposed method.

of all, the effect of our instance-aware contrastive learning scheme is analyzed and the corresponding result is shown in Table 4. As can be seen in Table 4, the performance for occluded human mesh reconstruction is significantly improved by leveraging the identity information into latent features for centers and their related joints. Furthermore, we can see that applying positional encoding to the center-aligned instance map is also useful for the model to discriminatively represent each center feature under severe occlusion cases. The effect of our instance-aware learning scheme is shown in Fig. 7. In the following, we check the performance variation by the change of the feature dimension for both center-aligned and joint-aligned instance maps is shown in Table 5. The best performance is achieved when the feature dimension is set to 64, thus this setting has been used as default for the performance evaluation of the proposed method. Note that the performance in all the settings is better than our baseline (i.e., the first row of Table 4). Based on ablation studies, we can conclude that the proposed instance-aware contrastive learning scheme is effective for occluded human mesh reconstruction under various real-world environments.

5. Conclusions

In this paper, we present a simple yet powerful method for occluded human mesh reconstruction, especially focusing on person-to-person occlusions. The core of the proposed method is to embed the identity information into latent features for body centers and joints of each person. To do this, we propose a novel instance-aware contrastive learning scheme. Experimental results on benchmark datasets show that the proposed method works reliably even under various person-to-person occlusions occurring in real-world environment.

Acknowledgments. This work was supported by Institute of Information Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00207, Immersive Media Research Laboratory).

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3686–3693, 2014. 6

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it SMPL: Automatic estimation of 3D human pose and shape from a single image. In *Proc. Eur. Conf. Comput. Vis.*, pages 561–578, 2016. 1, 2, 3, 5

[3] Junuk Cha, Muhammad Saqlain, GeonU Kim, Mingyu Shin, and Seungryul Baek. Multi-person 3D pose and shape estimation via inverse kinematics and refinement. In *Proc. Eur. Conf. Comput. Vis.*, pages 660–677, 2022. 3

[4] Hanbyel Cho, Yooshin Cho, Jaesung Ahn, and Junmo Kim. Implicit 3d human mesh recovery using consistency with pose and shape from unseen-view. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 21148–21158, 2023. 1, 3, 7

[5] Junhyeong Cho, Youwang Kim, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3D human mesh recovery with transformers. In *Proc. Eur. Conf. Comput. Vis.*, pages 342–359, 2022. 1, 3, 7

[6] 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 *Proc. Eur. Conf. Comput. Vis.*, pages 769–787, 2020. 3

[7] Hongsuk Choi, Gyeongsik Moon, JoonKyu Park, and Kyoung Mu Lee. Learning to estimate robust 3D human mesh from in-the-wild crowded scenes. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1475–1484, 2022. 1, 2, 3, 6, 7, 8

[8] Lin Geng Foo, Jia Gong, Hossein Rahmani, and Jun Liu. Distribution-aligned diffusion for human mesh recovery. In *Proc. Int. Conf. Comput. Vis.*, pages 9221–9232, 2023. 3, 6, 7

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 770–778, 2016. 7, 8

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7):1325–1339, 2013. 5, 7

[11] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5579–5588, 2020. 1, 3, 7

[12] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proc. Brit. Mach. Vis. Conf.*, pages 1–11, 2010. 6

[13] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proc. Int. Conf. Comput. Vis.*, pages 3334–3342, 2015. 6

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3d human model fitting towards in-the-wild 3d human pose estimation. In *Proc. Int. Conf. 3D Vis.*, pages 42–52, 2021. 6

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 7122–7131, 2018. 1, 2, 5

[16] Rawal Khirodkar, Shashank Tripathi, and Kris Kitani. Occluded human mesh recovery. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1715–1725, 2022. 1, 2, 3, 6, 7

[17] Jeonghwan Kim, Mi-Gyeong Gwon, Hyunwoo Park, Hyukmin Kwon, Gi-Mun Um, and Wonjun Kim. Sampling is Matter: Point-guided 3d human mesh reconstruction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12880–12889, 2023. 1, 3, 7

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proc. Int. Conf. Comput. Vis.*, pages 11127–11137, 2021. 1, 7

[19] 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 *Proc. Int. Conf. Comput. Vis.*, pages 2252–2261, 2019. 1, 2, 6, 7

[20] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 4501–4510, 2019. 3, 5

[21] Christoph Lassner, Javier Romero, Martin Kiefel, Federica Bogo, Michael J. Black, and Peter V. Gehler. Unite the people: Closing the loop between 3D and 2D human representations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 6050–6059, 2017. 1

[22] Haoyuan Li, Haoye Dong, Hanchao Jia, Dong Huang, Michael C Kampffmeyer, Liang Lin, and Xiaodan Liang. Coordinate transformer: Achieving single-stage multi-person mesh recovery from videos. In *Proc. Int. Conf. Comput. Vis.*, pages 8744–8753, 2023. 3, 6, 7

[23] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. CrowdPose: Efficient crowded scenes pose estimation and a new benchmark. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 10863–10872, 2019. 6

[24] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. HybrIK: A hybrid analytical-neural inverse kinematics solution for 3D human pose and shape estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 3383–3393, 2021. 1

[25] Jiefeng Li, Siyuan Bian, Qi Liu, Jiasheng Tang, Fan Wang, and Cewu Lu. Niki: Neural inverse kinematics with invertible neural networks for 3d human pose and shape estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 12933–12942, 2023. 1

[26] Zhihao Li, Jianzhuang Liu, Zhensong Zhang, Songcen Xu, and Youliang Yan. CLIFF: Carrying location information in full frames into human pose and shape estimation. In *Proc. Eur. Conf. Comput. Vis.*, pages 590–606, 2022. 3, 6, 7

[27] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In

Proc. IEEE Conf. Comput. Vis. Pattern Recog., pages 1954–1963, 2021. 1, 3, 7

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proc. Int. Conf. Comput. Vis.*, pages 12939–12948, 2021. 1, 3, 7

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 6, 7

[30] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 2980–2988, 2017. 5

[31] Qihao Liu, Yi Zhang, Song Bai, and Alan Yuille. Explicit occlusion reasoning for multi-person 3D human pose estimation. In *Proc. Eur. Conf. Comput. Vis.*, pages 497–517, 2022. 1, 3

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graph.*, 34(6):248:1–248:16, 2015. 1

[33] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018. 5

[34] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *Proc. Int. Conf. 3D Vis.*, pages 506–516, 2017. 5

[35] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3D pose estimation from monocular RGB. In *Proc. Int. Conf. 3D Vis.*, pages 120–130, 2018. 6

[36] Georgios Pavlakos, Luyang Zhu, Xiaowei Zhou, and Kostas Daniilidis. Learning to estimate 3D human pose and shape from a single color image. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 459–468, 2018. 7

[37] Karthik Shetty, Annette Birkhold, Srikrishna Jaganathan, Norbert Strobel, Markus Kowarschik, Andreas Maier, and Bernhard Egger. PLIKS: A pseudo-linear inverse kinematic solver for 3d human body estimation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 574–584, 2023. 1, 3

[38] Yu Sun, Qian Bao, Wu Liu, Yili Fu, Black Michael J., and Tao Mei. Monocular, one-stage, regression of multiple 3D people. In *Proc. Int. Conf. Comput. Vis.*, pages 11179–11188, 2021. 1, 3, 5, 6, 7, 8

[39] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. Putting people in their place: Monocular regression of 3D people in depth. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13243–13252, 2022. 1, 3

[40] Yu Sun, Lubing Xu, Qian Bao, Wu Liu, Wenpeng Gao, and Yili Fu. Learning monocular regression of 3d people in crowds via scene-aware blending and de-occlusion. *IEEE Trans. Multimedia*, 2023. 3

[41] 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 *Proc. Eur. Conf. Comput. Vis.*, pages 601–617, 2018. 6, 7, 8

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(10):3349–3364, 2021. 8

[43] Kaibing Yang, Renshu Gu, Maoyu Wang, Masahiro Toyoura, and Gang Xu. LASOR: Learning accurate 3D human pose and shape via synthetic occlusion-aware data and neural mesh rendering. *IEEE Trans. Image Process.*, 31:1938–1948, 2022. 3

[44] Yusuke Yoshiyasu. Deformable mesh transformer for 3d human mesh recovery. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 17006–17015, 2023. 1

[45] Wang Zeng, Sheng Jin, Wentao Liu, Chen Qian, Ping Luo, Wanli Ouyang, and Xiaogang Wang. Not all tokens are equal: Human-centric visual analysis via token clustering transformer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 11101–11111, 2022. 1

[46] Hongwen Zhang, Yating Tian, Xinchi Zhou, Wanli Ouyang, Yebin Liu, Limin Wang, and Zhenan Sun. PyMAF: 3d human pose and shape regression with pyramidal mesh alignment feedback loop. In *Proc. Int. Conf. Comput. Vis.*, pages 11446–11456, 2021. 1, 2, 6, 7

[47] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 546–556, 2021. 1, 3, 7

[48] Song-Hai Zhang, Ruilong Li, Xin Dong, Paul Rosin, Zixi Cai, Xi Han, Dingcheng Yang, Haozhi Huang, and Shi-Min Hu. Pose2Seg: Detection free human instance segmentation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 889–898, 2019. 6, 7

[49] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *Proc. Eur. Conf. Comput. Vis.*, pages 550–566, 2020. 3

[50] Ce Zheng, Xianpeng Liu, Guo-Jun Qi, and Chen Chen. Potter: Pooling attention transformer for efficient human mesh recovery. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1611–1620, 2023. 1

[51] Ce Zheng, Matias Mendieta, Taojiannan Yang, Guo-Jun Qi, and Chen Chen. FeatER: An efficient network for human reconstruction via feature map-based transformer. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 13945–13954, 2023. 1

[52] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. MonoCap: Monocular human motion capture using a CNN coupled with a geometric prior. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(4):901–914, 2018. 7