CenterHMR: Multi-Person Center-based Human Mesh Recovery

Yu Sun
Harbin Institute of Technology
yusun@stu.hit.edu.cn

Qian Bao
JD AI Research
baoqian@jd.com

Wu Liu
JD AI Research
liuwu@live.cn

Yili Fu
Harbin Institute of Technology
meylfu@hit.edu.cn

Michael J. Black
MPI for Intelligent Systems
black@tuebingen.mpg.de

Tao Mei
JD AI Research
tmei@live.com

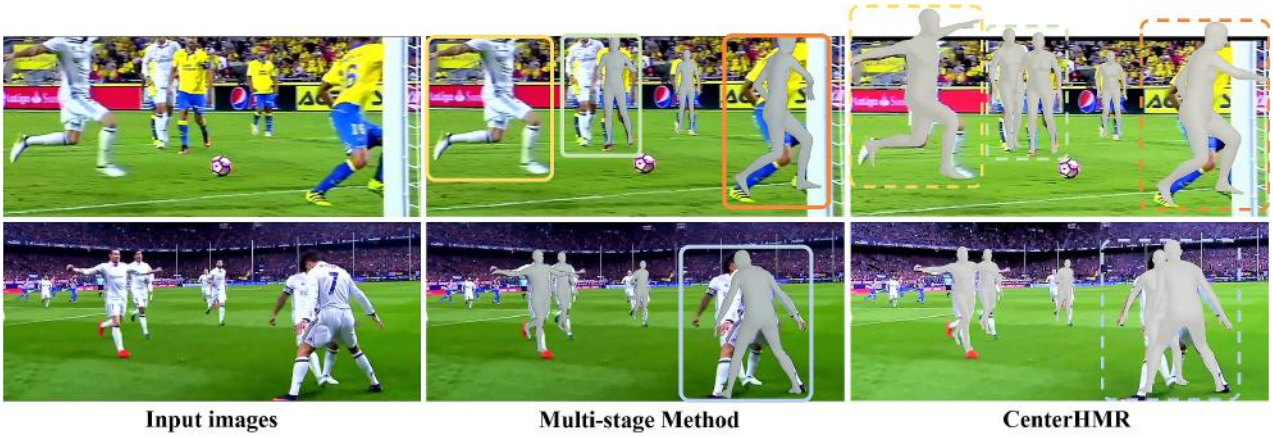

Figure 1. Given challenging multi-person images, the recent state-of-the-art human pose and shape estimation methods like VIBE [17] (middle) fail to deal with truncation (gold box), scene occlusion (orange box), and person-person occlusion (green and blue boxes). The architecture of such methods uses global features at the bounding box level that are ambiguous in these cases. To address this, we present a bottom-up single-shot network, the Center-based Human Mesh Recovery network (CenterHMR), which uses a pixel-level representation that increases robustness to the truncation and occlusion.

Abstract

This paper focuses on multi-person 3D mesh recovery from a single RGB image. Existing approaches predominantly follow a multi-stage pipeline, that detects bounding boxes and then regresses the body from bounding-box-level features. However, multi-person occlusion and truncation can make these features ambiguous, which results in the failure of recovery. To deal with this problem, we present a novel bottom-up single-shot method, named "Center-based Human Mesh Recovery network (CenterHMR)". The key idea is to develop an explicit center-based representation for bottom-up pixel-level estimation. Guided by the body centers, our model effectively locates every person and learns robust and discriminative features under occlusion. In an end-to-end manner, the model is trained to estimate multiple differentiable maps that contain the information of multi-person 3D body meshes and their locations. Furthermore, when encountering severe multi-person occlusion, the body centers may be very close or even overlapping. A collision-aware center representation is developed to ensure a distinguishable distance between body centers. Our proposed CenterHMR achieves state-of-the-art performance on four challenging multi-person/occlusion benchmarks (3DPW, CMU Panoptic, MuPoTs-3D, and 3DOH50K). Experiments on crowded/occluded datasets demonstrate the stability under various types of occlusion. Due to the concise bottom-up single-shot design, our released demo code[1] is the first open-source real-time (over 30 FPS) implementation of monocular multi-person 3D mesh recovery.

1. Introduction

Recently, great progress has been made on monocular 3D human pose and shape estimation, particularly in the case of a single person in the scene [3, 8, 15, 17, 18, 38, 40, 48]. However, as we progress towards more general cases, it is crucial to deal with the truncation, environmental occlusion, and person-person occlusion. Robustness to such occlusions is critical for real-world applications.

Existing methods [12, 17, 42, 43] follow a multi-stage design that equips the single-person pipeline with a 2D person detector to handle multi-person scenes. Generally, they first detect the person areas and then extract the bounding-

[1] https://github.com/Arthur151/CenterHMR

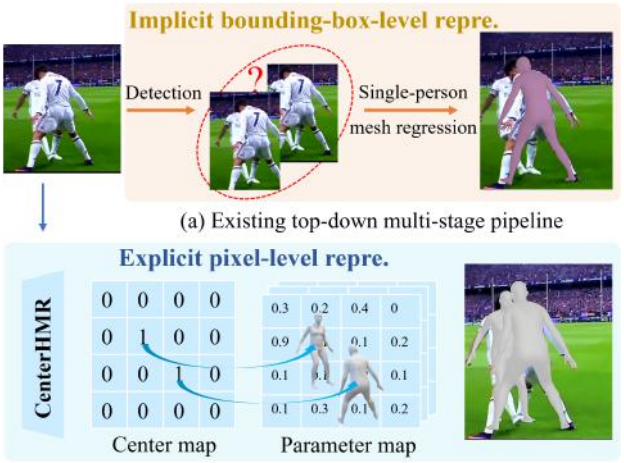

Figure 2. Existing multi-stage methods [12, 17] adopt a bounding-box-level representation, in which features are implicit, ambiguous, and inseparable in multi-person cases. In contrast, we develop an explicit pixel-level representation for fine-grained estimation.

box-level feature vectors from them, which are used to regress a single 3D human mesh [8, 15, 16, 17, 18, 19, 32, 37, 40, 48]. However, as shown in Figure 1, this strategy is prone to fail in cases of multi-person occlusion and truncation. For example, in Figure 2(a), when two people overlap, it is hard for the multi-stage method to estimate two diverse body meshes from two similar image patches. It is the ambiguity of the implicit bounding-box-level representation that leads to the failure in such inseparable multi-person cases.

For multi-person 2D pose estimation, this problem is well tackled via a subtle and effective bottom-up framework. The paradigm is to first detect all body joints and then assign them to different people by joint grouping. It is the pixel-level body joint representation that guarantees their strong performance in crowded scenes [4, 5, 33]. However, it is non-trivial to extend the bottom-up process beyond joints [12]. Unlike 2D pose estimation that estimates dozens of body joints, we need to predict a human body mesh with thousands of vertices, making it hard to follow the paradigm of body joint detection and grouping.

To address these problems, we propose CenterHMR, a simple bottom-up single-shot network, for monocular multi-person 3D mesh recovery. The key idea is to develop an explicit center-based representation for bottom-up pixel-level estimation. Instead of extracting the implicit bounding-box-level feature for single-person 3D mesh regression, CenterHMR directly estimates multiple maps and derives multi-person 3D mesh results in a bottom-up manner. As shown in Figure 2(b), CenterHMR predicts a Center map and a Parameter map, representing the 2D position of body center and the corresponding 3D body mesh parameter, respectively. Specifically, at each position, the Center map contains the confidence of it being a human body center, and the Parameter map contains the human mesh parameter vector assuming that it is a body center. The pixel-level center-based representation explicitly points out the target from the background/occlusion, to effectively learn from multi-person overlapping cases. During inference, we sample the mesh parameter vectors from the Parameter map according to the body center position contained in the Center map. Benefited from this explicit pixel-level representation, CenterHMR is trained in an end-to-end manner via a simple parameter sampling process. Finally, we put the sampled SMPL parameters into SMPL body model to derive multi-person 3D body meshes.

Moreover, considering that the body center of severe overlapping people may collide at the same 2D position, we further develop the center-based representation into a collision-aware version, CAR. The key idea is to construct a repulsion field of body centers, where close body centers are treated as positive charges that are pushed away by the mutual repulsion. In this way, the body centers of the overlapping people would be more distinguishable. Especially in the face of severe overlap, most of the human body is invisible. Mutual repulsion will automatically push the center to the visible body area, which makes the model tend to sample 3D mesh parameters estimated from the position centered on the visible body parts. This improves our robustness under heavy occlusion between people.

Compared with previous SOTA methods for multi-person [12, 42, 43] and single-person [17, 18, 44] 3D mesh recovery, CenterHMR achieves better results on four benchmarks, 3DPW [39], CMU Panoptic [13], MuPoTs-3D [27], and 3DOH50K [44]. Experiments on person occlusion datasets (Crowdpose [21] and 3DPW-PC, a person-occluded subset of 3DPW [39]) demonstrate the effectiveness of the proposed collision-aware representation (CAR) under person-person occlusion. To further evaluate it in general cases, we test CenterHMR on images from the Internet and web camera videos. CenterHMR achieves real-time performance with over 30 FPS on a 1070Ti.

In summary, the contributions are:

- A bottom-up single-shot network, CenterHMR, is proposed for monocular multi-person 3D mesh recovery.
- Our explicit center-based representation is novel and facilitates pixel-level human mesh estimation in an end-to-end manner.
- We develop an collision-aware representation to deal with the severe overlapping cases.
- State-of-the-art results are achieved on multiple benchmarks. Additionally, CenterHMR is the first open-source real-time method for multi-person 3D mesh recovery from monocular images.

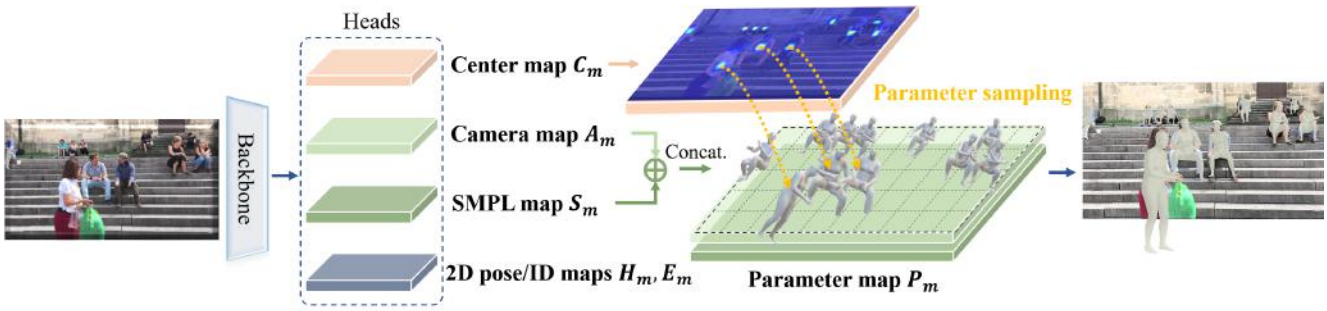

Figure 3. An overview of CenterHMR. Given an input image, CenterHMR predicts multiple maps: 1) the Center map predicts the probability of each position being a body center, 2) the Camera map and 3) SMPL map contain the camera and SMPL parameters of the person at each center, respectively, and 4) 2D pose/identity (ID) maps for promoting the representation learning. As the combination of the Camera map and SMPL map, the Parameter map contains the information of the predicted 3D body mesh and its location. Via the designed parameter sampling process, we can obtain the final 3D mesh results by parsing the Center map and sampling the Parameter map.

2. Related Work

Single-person 3D mesh recovery. Parametric human body models, like SMPL [25], have been widely adopted to encode the complex 3D human mesh into a low-dimensional parameter vector. Therefore, many methods tend to estimate the SMPL parameters instead of the 3D mesh, to reduce the complexity. Recently, impressive performance has been achieved for single-person scenes using various weak supervision signals, such as 2D pose [15, 37], semantic segmentation [40], geometric prior [15], motion dynamics [16], temporal coherence [17, 37], texture consistency [32], SMPLify [3] in the loop [18], etc. In this way, the available 2D/3D data is well explored to alleviate the lack of 3D data. However, all these methods adopt a bounding-box-level representation, which is implicit and ambiguous for the multi-person cases. For object occlusion, Zhang et al. [44] use a 2D UV map to represent a 3D human mesh. Considering that the object-occluded body parts are blank areas in the partial UV map, they propose to in-paint the partial UV map to make up the occluded information. However, in the case of person-person occlusion where one person's body parts are occluded by those of another, it is hard to generate the partial UV map.

Multi-person 3D pose estimation. The mainstream methods can be roughly divided into two categories, the multi-stage paradigm and the single-shot paradigm. Many top-down methods follow the design of the well-known Faster R-CNN [35], such as LCR-Net++ [36] and 3DMPPE [28]. Using the anchor-based feature proposal, they can directly estimate the target via regression. Other works explore the single-shot solution. They firstly estimate the body joint position in a heatmap manner and then associate the joints of each person via grouping. Mehta et al. [27] propose an occlusion-robust joint heatmap via making the redundant joint estimation at multiple positions. Benzine et al. [2] propose an anchor-based single-shot model, which directly estimates the 2D/3D pose results for each anchor position. For person-person occlusion, Zhen et al. [45] adopt the PAFs of OpenPose [4] to make part association. Benefiting from the explicit bottom-up joint-based representation, such as the volumetric heatmap, 3D pose estimation methods show impressive performance. Our proposed CenterHMR extends the end-to-end single-shot process beyond the body joints.

Multi-person 3D mesh recovery. There are a few recently proposed methods for multi-person 3D mesh recovery. Zanfir et al. [43] estimate the 3D mesh of each person from its intermediate 3D pose estimate. Zanfir et al. [42] further employ multiple scene constraints to optimize the multi-person 3D mesh results. Jiang et al. [12] propose a network for Coherent Reconstruction of Multiple Human (CRMH), which is built on Faster-RCNN [35]. They use the RoI-aligned feature of each person to predict the SMPL parameters. Additionally, they develop the interpolation and depth ordering loss to supervise the relative position between multiple people. All existing methods follow a multi-stage design. The complex multi-step process requires repeated feature extraction, which slows down the computational efficiency. The ambiguity of bounding-box-level feature makes them vulnerable to occlusion and truncation in multi-person cases. Instead, we propose a simple but effective method for end-to-end learning.

Pixel-level representations have proven useful in anchor-free detection methods, such as [6, 20]. They attempt to directly estimate the corner point of the bounding box in a heatmap manner. In this way, they can avoid the dense proposal of the anchor-based representation. Our method draws inspiration from them to develop pixel-level fine-grained representation. In contrast to the bounding box center used in [46], our body center is determined by the body joints, which is introduced in Sec. 3.3. Furthermore, we develop a collision-aware version of center representation to deal with the inherent center collision problem.

3. Method

3.1. Overview

The single-shot CenterHMR framework is illustrated in Figure 3. It adopts a simple multi-head design with a backbone and four head networks. Given a single RGB image as input, it outputs a Center map, Camera map, SMPL map, and 2D pose and identity (ID) maps, describing the detailed information of the estimated 3D human mesh. In the Center map, we predict the probability of each position being a human body center. At each position of the Camera/SMPL map, we predict the camera/SMPL parameters of the person that takes the position as the center. In addition, we predict the 2D pose heatmap and associative embedding as 2D pose and ID maps to promote representation learning. For simplicity, we combine the Camera map and SMPL map into the Parameter map. During inference, we utilize the 2D center coordinates parsed from the Center map to sample the corresponding parameter results from the Parameter map. Finally, we put the sampled parameters into the SMPL body model to generate the estimated 3D human meshes.

3.2. Basic Representations

We introduce the detailed representation of each map. Each output map is of size $n \times H \times W$, where n is the number of channels and $H = W = 64$.

Center map: $\boldsymbol{C_m} \in \mathbb{R}^{1\times H\times W}$ is a heatmap representing the 2D human body center in the image. Each human body center is represented as a Gaussian distribution in the Center map. For better representation learning, the Center map also integrates the scale information of the human body in the 2D image. Specifically, we calculate the Gaussian kernel size k of each person center in terms of its 2D body scale in the image. Given the diagonal length d_{bb} of the person bounding box and the width W of the Center map, the Gaussian kernel size is derived from

$$k = k_l + (\frac{d_{bb}}{\sqrt{2}W})^2 k_r, \tag{1}$$

where k_l is the minimum kernel size and k_r is the variation range of k. We set $k_l = 2$ and $k_r = 5$ by default.

Parameter map: $\boldsymbol{P_m} \in \mathbb{R}^{145\times H\times W}$ consists of two parts, the Camera map and SMPL map. Assuming that each location of these maps is the center of a human body, we estimate the corresponding 3D human body parameters. Specifically, we estimate the parameters of SMPL, which encode the 3D body mesh into a set of low-dimensional parameters. Estimating these SMPL parameters instead of the complex 3D mesh greatly reduces the complexity of our task. Additionally, following the previous works [15, 37], we employ a weak-perspective camera model to project the K 3D body joints $\boldsymbol{J^{3D}} \in \mathbb{R}^{K\times 3}$ of the estimated mesh back to the 2D image plane $\boldsymbol{J^{p2D}} \in \mathbb{R}^{K\times 2}$. This facilitates training the model with in-the-wild 2D pose datasets with varied imagery (e.g. COCO [23]), which helps with robustness and generalization.

Camera map: $\boldsymbol{A_m} \in \mathbb{R}^{3\times H\times W}$ contains the 3-dim camera parameters (s, t_x, t_y) which describe the 2D scale s and translation $\boldsymbol{t} = (t_x, t_y)$ of the people in the image. The scale s reflects the size and depth of the human body to some extent. t_x and t_y, ranging in $(-1, 1)$, reflect the normalized translation of the human body relative to the image center on the x and y axis, respectively. The mapping from $\boldsymbol{J^{3D}}$ to $\boldsymbol{J^{p2D}}$ can be derived as

$$\boldsymbol{J^{p2D}} = s\boldsymbol{J^{3D}} + \boldsymbol{t}. \tag{2}$$

The translation parameters allow more accurate position estimates than the Center map and we exploit these later for collision-aware repulsion.

SMPL map: $\boldsymbol{S_m} \in \mathbb{R}^{142\times H\times W}$ contains the 142-dim SMPL parameters, which describe the 3D pose and shape of the 3D human body mesh. SMPL establishes an efficient mapping from the pose $\boldsymbol{\theta}$ and shape $\boldsymbol{\beta}$ parameters to the human 3D body mesh $\boldsymbol{M} \in \mathbb{R}^{6890\times 3}$. The shape parameter $\boldsymbol{\beta} \in \mathbb{R}^{10}$ is the top-10 PCA weights of the SMPL statistical shape space. The pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{6\times 22}$ contain the 3D rotation of the 22 body joints in a 6D representation [47]. Instead of using the full 24 joints of the original SMPL model, we drop the last two hand joints. The first 3D joint rotation is the 3D body orientation in the camera coordinates, while the remainders are the relative 3D orientations of each body part with respect to its parent in a kinematic chain. The joints $\boldsymbol{J^{3D}}$ are derived via $\boldsymbol{P_{j3d}M}$ where $\boldsymbol{P_{j3d}} \in \mathbb{R}^{K\times 6890}$ is a sparse weight matrix that describes the linear mapping from 6890 vertices of the body mesh to the K body joints.

2D pose and ID maps: These are composed of two parts: a 2D pose heatmap $\boldsymbol{H_m} \in \mathbb{R}^{17\times H\times W}$, and an associative embedding [30] map $\boldsymbol{E_m} \in \mathbb{R}^{17\times H\times W}$. Here we intend to promote multi-task representation learning [41] via training with the related task of multi-person 2D pose estimation, supervised by the 2D pose map and person identity (ID) map. Each body joint has a heatmap to represent its 2D location and an embedding map to represent its person ID. $\boldsymbol{H_m}$ adopts the Gaussian distribution with constant kernel size 5 to represent the 2D joint locations. At each position of $\boldsymbol{E_m}$, we train the model to predict a unique feature vector as the ID embedding of the corresponding person.

3.3. CAR: Collision-Aware Center Representation

The entire framework is based on a concise center-based representation. It is crucial to define an explicit and robust body center so that the model can easily estimate the center location in various cases. Here we introduce the basic definition of the body center for the general case and its advanced version for severe occlusion.

Basic definition of the body center. Existing corner-based methods [6, 46] define the center of the bounding

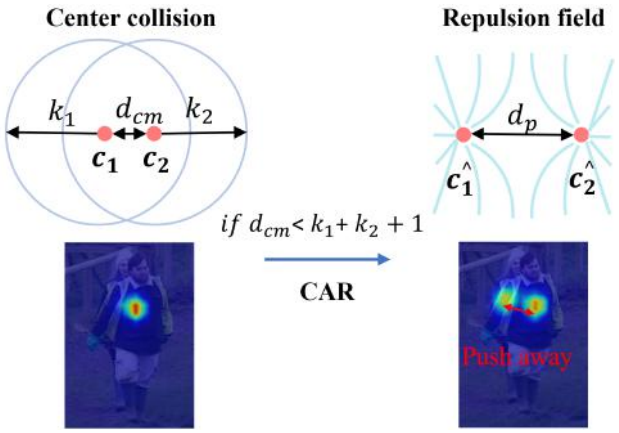

Figure 4. **Collision-Aware center Repulsion (CAR).** The body centers of overlapping people are treated as positive charges, which are pushed away if they are too close in the repulsion field.

box as the body center. This works well for general objects (such as a ball or bottle) that lack semantically meaningful keypoints. However, the center of the bounding box is not a meaningful point on the human body and may even fall outside the body area. For stable parameter sampling, we define an explicit body center. Each body center is calculated from the ground truth 2D pose. Considering that any single body joint may be occluded in general cases, we define the body center as the center of visible torso joints (neck, left/right shoulders, pelvis and left/right hips). When all torso joints are invisible, the center is simply determined by the average of the visible joints. In this way, the model is encouraged to estimate the results from the visible parts.

However, in cases of severely overlapping people, the body center of the people might be very close or even at the same location on $\boldsymbol{C_m}$. This center collision problem makes the center ambiguous and hard to identify in crowded cases. To tackle this, we develop a more robust representation to deal with person-person occlusion. To alleviate the ambiguity, the center points of overlapping people should be kept at a minimum distance to ensure that they can be well distinguished. Additionally, to avoid sampling multiple parameters for the same person, the network should assign a unique and explicit center for each person.

Based on these principles, we develop a novel **Collision-Aware center Representation (CAR)**. To ensure that the body centers are far enough from each other, we construct a repulsion field. In this field, each body center is treated as a positive charge, whose radius of repulsion is equal to its Gaussian kernel size derived by Eq. (1). In this way, the closer the body centers are, the greater the mutual repulsion and the further they will be pushed apart. Figure 4 illustrates the principle of CAR. Let $\boldsymbol{c_1}, \boldsymbol{c_2}$ be body centers of two overlapping people. If their Euclidean distance d_{cm} and Gaussian kernel sizes k_1, k_2 satisfy $d_{cm} < k_1 + k_2 + 1$, the repulsion is triggered to push the close centers away via

$$\begin{aligned} \hat{\boldsymbol{c_1}} &= \boldsymbol{c_1} - \gamma d_p, \hat{\boldsymbol{c_2}} = \boldsymbol{c_2} + \gamma d_p, \\ d_p &= \frac{k_1 + k_2 + 1 - d_{cm}}{d_{cm}} (\boldsymbol{c_1} - \boldsymbol{c_2}), \end{aligned} \tag{3}$$

where γ is an intensity coefficient to adjust the strength.

During training, we use CAR to push apart close body centers to supervise the Center map. In this way, the model is encouraged to estimate the centers that maintain a distinguishable distance. For the Center map, it helps the model to effectively locate the occluded person. For the Parameter map, sampling the parameters from these shifted locations enables the model to extract diverse and individual features for each person. The model trained with CAR is more suitable for crowded scenes with significant person-person occlusion such as train stations, canteens, etc.

3.4. Parameter Sampling

To estimate 3D human body meshes from the image, we need to first parse the 2D body center coordinates $\boldsymbol{c} \in \mathbb{R}^{K \times 2}$ from $\boldsymbol{C_m}$ and then use them to sample the $\boldsymbol{P_m}$ to obtain the SMPL parameters. In this section, we introduce the process of the center parsing, matching and sampling.

$\boldsymbol{C_m}$ is a probability map whose local maxima are the body centers. The local maxima are derived via $Mp(\boldsymbol{C_m}) \wedge \boldsymbol{C_m}$ where Mp is the max pooling operation and $\wedge$ is the logical conjunction operation. Let $\boldsymbol{c}$ be the 2D coordinates of a local maximum with confidence value larger than a threshold t_c. We rank the confidence value at each $\boldsymbol{c}$ and take the top N as the final centers. During inference, we directly sample the parameter results from $\boldsymbol{P_m}$ at $\boldsymbol{c}$. During training, the estimated $\boldsymbol{c}$ are matched with the nearest ground truth body center according to the L_2 distance.

Additionally, we approximate the depth order between multiple people using the center confidence from $\boldsymbol{C_m}$ and the 2D body scale s of the camera parameters from $\boldsymbol{A_m}$. For people of different s, the one with a larger s is in the front. For people of similar s, the person with a higher center confidence is considered to be in the front.

3.5. Loss functions

To supervise CenterHMR, we develop individual loss functions for different maps. In total, CenterHMR is supervised by the weighted sum of the center loss L_c, parameter loss L_p, and 2D pose loss L_{2D}.

Center loss. L_c encourages a high confidence value at the body center $\boldsymbol{c}$ of the Center map $\boldsymbol{C_m}$ and low confidence elsewhere. To deal with the imbalance between the center location and the non-center locations in $\boldsymbol{C_m}$, we train the Center map based on the focal loss [22]. Given the predicted

Center map $\boldsymbol{C_m^p}$ and the ground truth $\boldsymbol{C_m^{gt}}$, L_c is defined as

$$\begin{aligned} L_c &= -\frac{L_{pos}+L_{neg}}{\sum I_{pos}} w_c, \\ L_{neg} &= log(1-\boldsymbol{C_m^p})(\boldsymbol{C_m^p})^4(1-I_{pos}), \\ L_{pos} &= log(\boldsymbol{C_m^p})(1-\boldsymbol{C_m^p})^2 I_{pos}, I_{pos} = \boldsymbol{C_m^{gt}} \geq 1, \end{aligned} \tag{4}$$

where I_{pos} is a binary matrix with a positive value at the body center location, and w_c is the loss weight.

Parameter loss. As we introduced in Sec. 3.4, the parameter sampling process matches each ground truth body with a predicted parameter result for supervision. The parameter loss is derived as

$$\begin{aligned} L_p = w_{pose}L_{pose} + w_{shape}L_{shape} + w_{j3d}L_{j3d} \\ + w_{paj3d}L_{paj3d} + w_{pj2d}L_{pj2d} + w_{prior}L_{prior}. \end{aligned} \tag{5}$$

L_{pose} is the L_2 loss of the pose parameter in the 3×3 rotation matrix format. L_{shape} is the L_2 loss of the shape parameters. L_{j3d} is the L_2 loss of the 3D joints $\boldsymbol{J^{3D}}$ regressed from the body mesh $\boldsymbol{M}$. L_{paj3d} is the L_2 loss of the 3D joints $\boldsymbol{J^{3D}}$ after Procrustes alignment. L_{pj2d} is the L_2 loss of the projected 2D joints $\boldsymbol{J^{p2D}}$ via Eq. (2). L_{prior} is the Mixture Gaussian prior loss of the SMPL parameters adopted from [3, 25] for supervising the plausibility of the 3D joint rotations and body shape. Finally, $w_{(.)}$ denotes the corresponding loss weights.

2D pose estimation loss. Following [5] and [30], the 2D pose loss is defined as

$$L_{2D} = w_{hm}L_{hm} + w_{ae}L_{ae}, \tag{6}$$

where L_{hm} is the L_2 loss of the heatmap $\boldsymbol{H_m}$, representing the error of 2D body joint detection, and L_{ae} is the joint grouping loss of the ID map $\boldsymbol{E_m}$. Each joint heatmap corresponds to an ID map. We sample the ID embedding vector of each joint from $\boldsymbol{E_m}$ at the 2D joint position estimated in $\boldsymbol{H_m}$. L_{ae} supervises the model to shorten the L_2 distance of the ID embedding vectors within the same person and to expand the distance across different people.

4. Experiments

In this section, we first introduce the implementation details and experimental settings. Then we conduct experiments on multi-/single-person datasets to compare CenterHMR with state-of-the-art approaches. We further report results on occlusion datasets and finally conduct ablation experiments to show the effectiveness of our contributions.

4.1. Implementation Details

Network Architecture. HRNet-32 [5] is adopted as the backbone for its impressive performance in crowded scenes. Through its powerful recurrent multi-resolution aggregation, a feature vector $\boldsymbol{f_b} \in \mathbb{R}^{32\times H_b \times W_b}$ is extracted from a single RGB image. Also, we adopt the CoordConv [24] to enhance the spatial information at each center. Therefore, the backbone feature $\boldsymbol{f} \in \mathbb{R}^{34\times H_b \times W_b}$ is the combination of a coordinate index map $\boldsymbol{ci} \in \mathbb{R}^{2\times H_b \times W_b}$ and $\boldsymbol{f_b}$. Next, from $\boldsymbol{f}$, four head networks are developed to estimate the Center, Camera, SMPL, and 2D pose/ID maps. Each head network is composed of two basic ResNet blocks [9] with batch normalization. A softmax layer is added at the end of the Center map branch for normalization.

Setting Details. The input images are resized to 512×512, keeping the same aspect ratio and padding with zeros. The size of the backbone feature is $H_b = W_b = 128$. The maximum person number $N = 15$ could be determined by the certain situation. The loss weights are set to $w_c = 200, w_{j3d} = 360, w_{paj3d} = 400, w_{pj2d} = 420, w_{pose} = 60, w_{shape} = 1, w_{prior} = 1.6, w_{hm} = 60, w_{ae} = 500$ to ensure that the weighted loss items are in the same magnitude. The threshold of Center map is $t_c = 0.25$ and the intensity coefficient of CAR is $\gamma = 0.2$.

Evaluation benchmarks and protocols. 3DPW [39] is a large-scale in-the-wild 3D dataset containing multi-/single-person videos with abundant 2D/3D annotations, such as 2D pose, 3D pose, SMPL parameters, human 3D mesh, etc. We adopt three evaluation protocols to make a fair comparison. Following the ECCV 2020 3DPW Challenge, *Protocol 1* uses the entire dataset for evaluation without any ground truth for fine-tuning or human image cropping. Following VIBE [17], *Protocol 2* evaluates the model on the test set only. Different from *Protocol 2*, *Protocol 3* uses the 3DPW training set for fine-tuning. **CMU Panoptic [13]** and **MuPoTs-3D [27]** are multi-person 3D pose benchmarks. They are used for evaluation using the protocol provided by CRMH [12].

Occlusion datasets. 3DOH50K [44] is a single-person object-occluded dataset providing various 2D/3D annotations, such as 2D pose, 3D pose, and SMPL parameters. **Crowdpose** [21] is a crowded-person dataset with 2D pose annotation. We evaluate on its test set for the crowd-scene evaluation. Additionally, we employ subsets of 3DPW [39], the person-person occluded **3DPW-PC**, the object-occluded **3DPW-OC** and the non-occluded/truncated **3DPW-NC** for occlusion evaluation. Please refer to the Supplementary Material for more details.

Training datasets. We use three indoor single-person 3D pose datasets (Human3.6M [11], MPI-INF-3DHP [26], and MoVi [7]) and a synthetic multi-person 3D pose dataset (MuCo-3DHP [26]) for training. Following previous methods [12, 17], we also train with multiple in-the-wild 2D pose datasets, such as COCO [23], for better generalization. Following SPIN [18], CenterHMR uses the pseudo-3D-label of part 2D pose datasets for training.

Evaluation metrics. We adopt per-vertex error (PVE) to evaluate the 3D surface error. To evaluate the 3D pose accuracy, we employ mean per joint position error (MPJPE), Procrustes-aligned MPJPE (PA-MPJPE), per-

Table 1. Comparisons to the SOTA methods on 3DPW following *Protocol 1* (without using any ground truth during inference). * means training with extra data from [14].

Method	Pub.	MPJPE (↓)	PA-MPJPE (↓)	PCK(↑)	AUC(↑)	MPJAE(↓)	PA-MPJAE(↓)	PVE(↓)
OpenPose + SPIN [18]	ICCV19	95.8	66.4	33.3	55.0	23.9	24.4	-
YOLO + VIBE [17]	CVPR20	94.7	66.1	33.9	56.6	25.2	20.46	112.7
CRMH [12]	CVPR20	105.9	71.8	28.5	51.4	26.4	22.0	120.9
CenterHMR	-	**81.8**	**58.6**	**37.3**	**59.9**	**20.8**	**19.1**	**96.0**
CenterHMR*	-	**81.6**	**58.5**	**37.2**	**60.2**	**20.5**	**19.0**	**95.2**

Table 2. Comparisons to the SOTA methods on 3DPW following VIBE [17], using *Protocol 2* (on the test set only) (the upper rows) and *Protocol 3* (fine-tuned on the training set) (the lower rows). The comparison results are obtained from VIBE [17].

Method	MPJPE (↓)	PA-MPJPE (↓)	PVE(↓)
HMR [15]	130.0	76.7	-
Kanazawa et al. [16]	116.5	72.6	139.3
Arnab et al. [1]	-	72.2	-
GCMR [19]	-	70.2	-
DSD-SATN [37]	-	69.5	-
SPIN [18]	96.9	59.2	116.4
I2L-MeshNet [29]	93.2	58.6	-
VIBE [17]	93.5	56.5	113.4
CenterHMR	**85.5**	**53.2**	**103.0**
VIBE [17]+3DPW	82.9	51.9	99.1
CenterHMR+3DPW	**76.0**	**46.7**	**92.4**

centage of correct keypoints (PCK), area under the PCK-threshold curve (AUC), mean per joint angle error (MPJAE), and Procrustes-aligned MPJAE (PA-MPJAE). Additionally, average precision ($AP^{0.5}$) is used to evaluate 2D multi-person pose accuracy of the back-projected $\boldsymbol{J^{p2D}}$.

4.2. Comparisons to the State-of-the-Art

In-the-wild multi-/single-person dataset, 3DPW. We follow *Protocol 1* from the 3DPW Challenge to evaluate on the entire 3DPW dataset without using any ground truth during inference, especially the ground truth bounding box. For a fair comparison, the single-person methods [17, 18] are equipped with a human detector (OpenPose [4] or YOLO [34]). Tab. 1 compares CenterHMR to the state-of-the-art (SOTA) methods. The results of OpenPose + SPIN are obtained from [10]. The results of YOLO + VIBE are obtained using their officially released code, which already contains the YOLO part for human detection. In Tab. 1, we observe that CenterHMR significant improves all evaluation metrics, especially in MPJPE, PA-MPJPE, and PVE. These results demonstrate the robustness of the proposed center-based representation in generalizing to the in-the-wild multi-/single-person scenes. Results of training with extra data from [14] show that the performance may be further improved via adding the training data.

Next, to avoid the influence of detection during evaluation, we use *Protocol 2* from VIBE [17], which evaluates on the test set using the bounding box provided by VIBE [17] for image cropping. Furthermore, to test the domain adaptation ability, we follow *Protocol 3* from VIBE [17] to fine-tune our model on the training set and then evaluate it on the test set. In Tab. 2, CenterHMR outperforms all existing methods that adopt a bounding-box-level representation in both protocols.

Table 3. Comparisons to the SOTA methods on CMU Panoptic [13] benchmark. The evaluation metric is MPJPE (lower is better) after centering the root joint. All methods are directly evaluated without any fine-tuning.

Method	Haggling	Mafia	Ultim.	Pizza	Mean
Zanfir et. al. [43]	141.4	152.3	145.0	162.5	150.3
MSC [42]	140.0	165.9	150.7	156.0	153.4
CRMH [12]	129.6	133.5	153.0	156.7	143.2
CenterHMR	**110.0**	**123.3**	**139.8**	**135.4**	**127.1**

Table 4. Comparisons to the SOTA methods on MuPoTs-3D [27]. The evaluation metric is 3D PCK (higher is better). The comparison results are obtained from CRMH [12].

Method	All	Matched
OpenPose+SMPLify-X [31]	62.84	68.04
OpenPose+HMR [15]	66.09	70.90
CRMH [12]	69.12	72.22
CenterHMR	**69.90**	**74.60**

Table 5. Run-time comparisons on 1070Ti. HRNet-32 [5] is used as the default backbone. * is using ResNet-50 [9] as backbone.

Method	VIBE [17]	CRMH [12]	Ours	Ours*
FPS	10.9	14.1	20.8	**30.9**

Multi-person benchmarks. For a comprehensive comparison, we evaluate CenterHMR on the multi-person benchmarks, CMU Panoptic [13] and MuPoTs-3D [27], following the evaluation protocol of CRMH [12]. As shown in Tab. 3 and 4, CenterHMR outperforms the existing multi-stage methods [12, 42, 43], demonstrating the advantage of our novel single-shot design.

Runtime. To validate the computational efficiency, we compare CenterHMR with the existing SOTA methods in processing videos captured by a web camera. As shown in Tab. 5, CenterHMR achieves real-time performance (30.9 fps using ResNet-50 and 20.8 fps using HRNet-32), significantly faster than the competing methods. Additionally, compared with the multi-stage methods [12, 17], CenterHMR's processing time is roughly constant regardless of the number of people.

4.3. Experiments on occlusion datasets

To validate the stability under occlusion, we evaluate CenterHMR on multiple occlusion datasets. Firstly, on the **person-occluded** 3DPW-PC and Crowdpose [21], results

Figure 5. Ablation study (left) and qualitative results (right) of the proposed CAR on Crowdpose.

Table 6. Comparisons to the SOTA methods on the person-occluded (3DPW-PC), object-occluded (3DPW-OC) and non-occluded/truncated (3DPW-NC) subsets of 3DPW. The evaluation metric is PA-MPJPE (lower is better).

Method	3DPW-PC	3DPW-NC	3DPW-OC
CRMH [12]	103.5	68.7	78.9
VIBE [17]	103.9	57.3	65.9
CenterHMR	**80.1**	**53.7**	**65.5**
CenterHMR+CAR	**76.9**	54.8	66.0

Table 7. Comparisons to the SOTA methods on Crowdpose [21] benchmark. The evaluation metric is $AP^{0.5}$.

Split	CRMH [12]	CenterHMR	CenterHMR+CAR
Test	33.9	48.1	**51.7**
Validation	32.9	48.6	**51.8**

in Tab. 6 (second column) and 7 along with some qualitative examples in Figure 5 show that CenterHMR outperforms previous SOTA methods [12, 17]. This suggests that the pixel-level representation is important for improving the performance under person-person occlusion. Except for the quantitative results of the proposed CAR, we also provide the qualitative ablation study in Figure 5. Both results show that adding the proposed CAR further improves the performance, which demonstrates that CAR effectively tackles the center collision problem in crowded scenes. Finally, on the **object-occluded** datasets, 3DOH50K [44] and 3DPW-OC, CenterHMR also achieves SOTA results in Tab. 8. These results demonstrate that the fine-grained pixel-level representation is beneficial for dealing with various occlusion cases.

4.4. Ablation Study

Different subsets of 3DPW. To determine the source of our performance advantage, we conduct an ablation study on 3DPW subsets that contain different scenes. Results in Tab. 6 show that, compared with previous SOTA methods [12, 17], our main gains come from the person-occluded and the non-occluded/truncated cases. Our pixel-level representation helps to effectively learn from in-the-wild multi-person cases via explicitly directing the target from the background/occlusion. Further, in Tab. 6, we observe varying degrees of performance degradation by adding the CAR in these cases without strong person-person occlusion. The reason is probably that pushing away the body centers affects the consistency of the center-based representation.

Table 8. Comparisons to the SOTA methods on occlusion datasets, 3DOH50K and 3DPW-OC. The evaluation metric is PA-MPJPE. The compared results are obtained from [44].

Method	3DOH50K	3DPW-OC
HMR [15]	83.2	103.8
GCMR [19]	76.3	104.8
SPIN [19]	67.5	95.4
Zhang et al. [44]	58.5	72.2
CRMH [12]	56.9	78.9
CenterHMR	**36.9**	**65.5**

Table 9. Ablation study of CAR with diverse repulsion coeff. γ on person-occluded 3DPW-PC.

Method	MPJPE(↓)	PA-MPJPE(↓)
w/o CAR	103.7	80.1
0.1	**97.2**	78.1
0.2	97.6	**76.9**
0.3	97.8	79.5
0.4	99.5	80.1

Intensity coefficient γ of the CAR. To figure out the proper setting of the intensity coefficient γ, we conduct a further ablation study on 3DPW-PC. Results in Tab. 9 show that if γ is too large, the 3D pose accuracy on 3DPW-PC decreases. Therefore, we adopt $\gamma = 0.2$ to balance between the distinguishable center and consistent representation learning. More ablation studies of the network architecture are included in Supplementary Material.

5. Conclusion

We introduce a novel single-shot network, CenterHMR, for monocular multi-person 3D mesh recovery. For pixel-level estimation, we propose an explicit center-based representation and further develop with a collision-aware version, CAR, which enables robust prediction under the occlusion. CenterHMR is the first single-shot regression model that achieves state-of-the-art performance on several benchmarks as well as real-time inference speed. For the community, we believe that it is a simple but effective baseline for many other 3D multi-person tasks, such as depth estimation, tracking, and interaction modeling.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 7

[2] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. PandaNet: Anchor-based single-shot multi-person 3D pose estimation. In *CVPR*, 2020. 3

[3] 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 *ECCV*, 2016. 1, 3, 6

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2D pose estimation using part affinity fields. In *CVPR*, 2017. 2, 3, 7

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S. Huang, and Lei Zhang. HigherHRNet: Scale-aware representation learning for bottom-up human pose estimation. In *CVPR*, 2020. 2, 6, 7

[6] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *ICCV*, 2019. 3, 4

[7] Saeed Ghorbani, Kimia Mahdaviani, Anne Thaler, Konrad Kording, Douglas James Cook, Gunnar Blohm, and Nikolaus F Troje. MoVi: A large multipurpose motion and video dataset. *arXiv*, 2020. 6

[8] Riza Alp Guler and Iasonas Kokkinos. HoloPose: Holistic 3D human reconstruction in-the-wild. In *CVPR*, 2019. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 7

[10] Kissos Imry, Fritz Lior, Goldman Matan, Meir Omer, Oks Eduard, and Kliger Mark. Beyond weak perspective for monocular 3D human pose estimation. In *ECCVW*, 2020. 7

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

[12] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[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 *ICCV*, 2015. 2, 6, 7

[14] Hanbyul Joo, Natalia Neverova, and Andrea Vedaldi. Exemplar fine-tuning for 3D human pose fitting towards in-the-wild 3D human pose estimation. In *ECCV*, 2020. 7

[15] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 1, 2, 3, 4, 7, 8

[16] Angjoo Kanazawa, Jason Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 2, 3, 7

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. VIBE: Video inference for human body pose and shape estimation. In *CVPR*, 2020. 1, 2, 3, 6, 7, 8

[18] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 1, 2, 3, 6, 7

[19] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 2, 7, 8

[20] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 3

[21] 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 *CVPR*, 2019. 2, 6, 7, 8

[22] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

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

[24] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *NeurIPS*, 2018. 6

[25] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 3, 6

[26] 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 *3DV*, 2017. 6

[27] 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 *3DV*, 2018. 2, 3, 6, 7

[28] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. In *CVPR*, 2019. 3

[29] Gyeongsik Moon and Kyoung Mu Lee. I2L-MeshNet: Image-to-lixel prediction network for accurate 3D human

pose and mesh estimation from a single RGB image. In *ECCV*, 2020. 7

[30] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 4, 6

[31] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *CVPR*, 2019. 7

[32] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 2, 3

[33] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *CVPR*, 2016. 2

[34] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 7

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[36] Grégory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2D and 3D pose detection in natural images. *TPAMI*, 2019. 3

[37] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *ICCV*, 2019. 2, 3, 4, 7

[38] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 1

[39] Timo von Marcard, Roberto Henschel, Michael Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using imus and a moving camera. In *ECCV*, 2018. 2, 6

[40] Yuanlu Xu, Song-Chun Zhu, and Tony Tung. DenseRaC: Joint 3D pose and shape estimation by dense render-and-compare. In *ICCV*, 2019. 1, 2, 3

[41] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 4

[42] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *CVPR*, 2018. 1, 2, 3, 7

[43] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3D sensing of multiple people in natural images. In *NeurIPS*, 2018. 1, 2, 3, 7

[44] Tianshu Zhang, Buzhen Huang, and Yangang Wang. Object-occluded human shape and pose estimation from a single color image. In *CVPR*, 2020. 2, 3, 6, 8

[45] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. SMAP: Single-shot multi-person absolute 3D pose estimation. *ECCV*, 2020. 3

[46] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv*, 2019. 3, 4

[47] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 4

[48] Hao Zhu, Xinxin Zuo, Sen Wang, Xun Cao, and Ruigang Yang. Detailed human shape estimation from a single image by hierarchical mesh deformation. In *CVPR*, 2019. 1, 2