Monocular, One-stage, Regression of Multiple 3D People

Yu Sun[1*]   Qian Bao[2]   Wu Liu[2†]   Yili Fu[1†]   Michael J. Black[3]   Tao Mei[2]

[1]Harbin Institute of Technology   [2]JD AI Research

[3]Max Planck Institute for Intelligent Systems, Tübingen, Germany

yusun@stu.hit.edu.cn, baoqian@jd.com, liuwu@live.cn, meylfu@hit.edu.cn

black@tuebingen.mpg.de, tmei@live.com

Abstract

This paper focuses on the regression of multiple 3D people from a single RGB image. Existing approaches predominantly follow a multi-stage pipeline that first detects people in bounding boxes and then independently regresses their 3D body meshes. In contrast, we propose to Regress all meshes in a One-stage fashion for Multiple 3D People (termed ROMP). The approach is conceptually simple, bounding box-free, and able to learn a per-pixel representation in an end-to-end manner. Our method simultaneously predicts a Body Center heatmap and a Mesh Parameter map, which can jointly describe the 3D body mesh on the pixel level. Through a body-center-guided sampling process, the body mesh parameters of all people in the image are easily extracted from the Mesh Parameter map. Equipped with such a fine-grained representation, our one-stage framework is free of the complex multi-stage process and more robust to occlusion. Compared with state-of-the-art methods, ROMP achieves superior performance on the challenging multi-person benchmarks, including 3DPW and CMU Panoptic. Experiments on crowded/occluded datasets demonstrate the robustness under various types of occlusion. The code, released at https://github.com/Arthur151/ROMP, *is the first real-time implementation of monocular multi-person 3D mesh regression.*

1. Introduction

Recently, great progress has been made in monocular 3D human pose and shape estimation, particularly in images with a single person [20, 22, 23, 48, 51]. However, for more general scenes with multiple people, it is crucial to deal with truncation by the image frame, person-person occlusion, and environmental occlusion. Robustness to such occlusions is critical for real-world applications.

Existing approaches [15, 22, 52, 53] follow a multi-

*This work was done when Yu Sun was an intern at JD AI Research.
†Corresponding author.

(a) Challenges in the multi-person scene

(b) Existing multi-stage pipeline
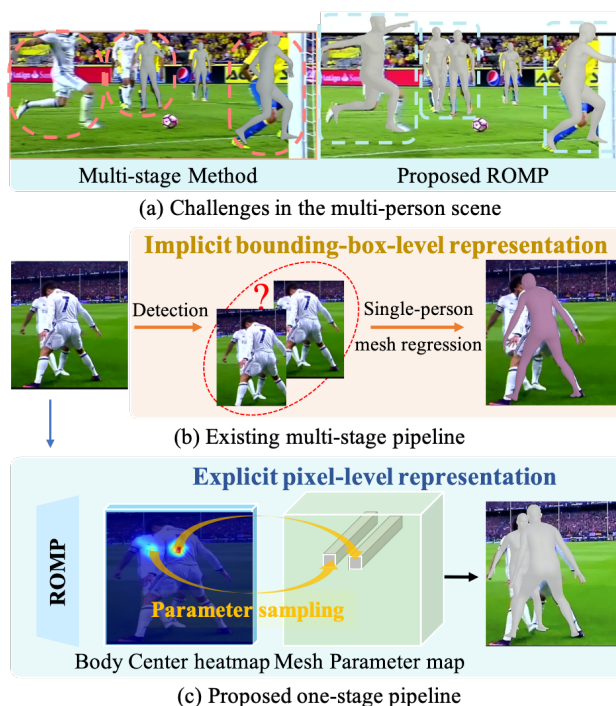
(c) Proposed one-stage pipeline

Figure 1. Given a challenging multi-person image like (a), the state-of-the-art approaches, e.g., VIBE [22] (left), fail to deal with truncation, scene occlusion, and person-person occlusion. The reason lies in the multi-stage design (b), where the bounding-box-level features are often implicit, ambiguous, and inseparable in multi-person cases. We propose to regress all meshes in one single stage for multiple 3D people. Specifically, we develop an explicit pixel-level representation (c) for fine-grained one-stage estimation that increases robustness to truncation and occlusion while significantly reducing computational complexity.

stage design that equips the single-person method with a 2D person detector to handle multi-person scenes. Generally, they first detect regions with people and then extract the bounding-box-level features, which are used to regress each single 3D human mesh [11, 20, 21, 22, 23, 24, 39, 47, 51, 60]. However, as shown in Fig. 1, this strategy is prone

to fail in cases of multi-person occlusion and truncation. Specifically, as shown in Fig. 1(b), when two people overlap, it is hard for the multi-stage method to estimate diverse body meshes from similar image patches. The ambiguity of the implicit bounding-box-level representation results in failure for such inseparable multi-person cases.

For multi-person 2D pose estimation, this problem has been tackled via a subtle and effective bottom-up framework. The paradigm first detects all body joints and then assigns them to different people by grouping body joints. This pixel-level body-joint representation enables their impressive performance in crowded scenes [7, 8, 40]. However, it is non-trivial to extend this bottom-up one-stage process beyond joints [15]. Unlike 2D pose estimation, which predicts dozens of body joints, we need to regress a human body mesh with thousands of vertices, making it hard to follow the paradigm of body joint detection and grouping.

Instead, we introduce ROMP, a one-stage network for regressing multiple 3D people in a per-pixel prediction fashion. It directly estimates multiple differentiable maps from the whole image, from which we can easily parse out the 3D meshes of all people. Specifically, as shown in Fig. 1(c), ROMP predicts a Body Center heatmap and a Mesh Parameter map, representing the 2D position of the body center and the parameter vectors of the corresponding 3D body mesh, respectively. Via a simple parameter sampling process, we extract 3D body mesh parameter vectors of all people from the Mesh Parameter map at the body center locations described by the heatmap. Then we put the sampled mesh parameter vectors into the SMPL body model [33] to derive multi-person 3D meshes. Following the guidance of body centers during training, the ambiguity of the regression target is greatly alleviated in crowded multi-person scenes. Additionally, in contrast to the local, bounding-box-level, features learned by traditional methods, end-to-end learning from the whole image forces the model to learn appropriate features from the holistic scene to predict bodies with occlusion. This holistic approach captures the complexity of real-world scenes, enabling the generalization and robustness to complex multi-person cases.

Moreover, since the body centers of severely overlapping people may collide at the same 2D position, we further develop an advanced collision-aware representation (CAR). The key idea is to construct a repulsion field of body centers, where close body centers are analogous to positive charges that are pushed apart by mutual repulsion. In this way, the body centers of the overlapping people are more distinguishable. Especially in the case of severe overlap, most part of the human body is invisible. Mutual repulsion will push the center to the visible body area, meaning that the model tends to sample 3D mesh parameters estimated from the position centered on the visible body parts. It improves the robustness under heavy person-person occlusion.

Compared with previous state-of-the-art methods for multi-person [15, 52, 53] and single-person [22, 23] 3D mesh regression, ROMP achieves superior performance on challenging benchmarks, including 3DPW [49] and CMU Panoptic [18]. Experiments on person-person occlusion datasets (Crowdpose [28] and 3DPW-PC, a person-occluded subset of 3DPW [49]) demonstrate the effectiveness of the proposed CAR under person-person occlusion. To further evaluate it in general cases, we test ROMP on images from the Internet and web camera videos. With the same backbone as the multi-stage counterparts, ROMP runs in real-time (over 30 FPS) on a 1070Ti GPU.

In summary, the contributions are: (1) To the best of our knowledge, ROMP is the first one-stage method for monocular multi-person 3D mesh regression, along with an open-source real-time implementation. Its simple yet effective framework leads to superior accuracy and efficiency. (2) The proposed explicit body-center-guided representation facilitates the pixel-level human mesh regression in an end-to-end manner. (3) We develop a collision-aware representation to deal with cases of severe overlap.

2. Related Work

Single-person 3D mesh regression. Parametric human body models, e.g., SMPL [33], have been widely adopted to encode the complex 3D human mesh into a low-dimensional parameter vector, which can be regressed from images [32]. Impressive performance has been achieved using various weak supervision signals, such as 2D pose [9, 20, 25], semantic segmentation [51], geometric priors and discriminative training [20], motion dynamics [21], temporal coherence [22, 47], texture consistency [39], optimization [6] in the loop [23], etc. However, all these methods adopt a bounding-box-level representation, which is implicit and ambiguous for multi-person/occlusion cases. To handle occluding objects, Zhang et al. [55] use a 2D UV map to represent a 3D human mesh. Considering the object-occluded body parts as blank areas in the partial UV map, they 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. Mainstream methods can be divided into two categories: the one-stage and the multi-stage. Many multi-stage methods follow the top-down design of Faster R-CNN [42], such as LCR-Net++ [43] and 3DMPPE [36]. From anchor-based feature proposals, they estimate the target via regression. Other works explore a one-stage solution that reasons about all people in a single forward pass. They estimate all the body joint positions and then group them into individuals. Mehta et al. [35] propose occlusion-robust pose-maps and exploit the body part association to avoid the bounding box predic-
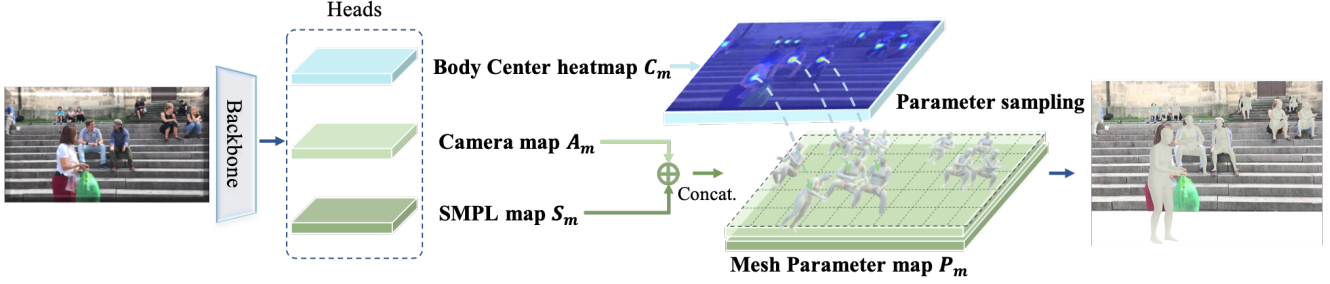
Figure 2. An overview of ROMP. Given an input image, ROMP predicts multiple maps: 1) the Body Center heatmap predicts the probability of each position being a body center, 2) the Camera map and 3) SMPL map contain the camera and SMPL parameters [33] of the person at each center, respectively. As the concatenation of the Camera map and SMPL map, the Mesh 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 Body Center heatmap and sampling the Mesh Parameter map.

tion. PandaNet [5] is an anchor-based one-stage model that relies on a huge number of pre-defined anchor predictions and the positive anchor selection. To handle person-person occlusion, Zhen et al. [56] extend part affinity fields [7] to be depth-aware. ROMP extends the end-to-end one-stage process beyond joints through a concise body-center-guided pixel-level representation and does not require the part association or an enormous number of anchor predictions.

Multi-person 3D mesh regression. There are only a few approaches for multi-person 3D mesh regression. Zanfir et al. [53] estimate the 3D mesh of each person from its intermediate 3D pose estimation. Zanfir et al. [52] further employ multiple scene constraints to optimize the multi-person 3D mesh results. Jiang et al. [15] propose a network for Coherent Reconstruction of Multiple Humans (CRMH). Built on Faster-RCNN [42], they use the RoI-aligned feature of each person to predict the SMPL parameters. Additionally, they learn the relative positions between multiple people via an interpenetration and depth ordering loss. All these methods follow a multi-stage design. The complex multi-step process requires a repeated feature extraction, which is computationally expensive. Moreover, since they rely on detected bounding boxes, the ambiguity and the limited local view of the bounding-box-level features make it hard to effectively learn from person-person occlusion and truncation. Instead, our proposed one-stage method learns an explicit pixel-level representation with a holistic view, which significantly improves both accuracy and efficiency in multi-person in-the-wild scenes.

Pixel-level representations have proven useful in anchor-free detection, such as CornerNet [27], Center-Net [10, 57], and ExtremeNet [58]. They directly estimate the corner or center point of the bounding box in a heatmap manner, which avoids the dense anchor-based proposal. Inspired by these, we develop a pixel-level fine-grained representation for multi-person 3D meshes. Different from the bounding box center used in [57], our body center is deter-

mined by the body joints, as introduced in Sec. 3.3. A recent work, BMP [54], uses a multi-scale grid-level representation for multi-person 3D mesh recovery, which locates a target person at the center of the grid cell.[1] In contrast to these methods, ROMP adopts a concise body-center-based representation and further evolves it to a collision-aware version to deal with the inherent center collision problem.

Disambiguation is a key goal of ROMP, enabling it to deal with the crowded multi-person scenes. Related techniques have been studied in many other fields. For instance segmentation, Adaptis [45] separately learns the segmentation mask of each instance selected by the guide point. To alleviate the ambiguity between embeddings of similar samples, associate embedding [38], triplet loss [44], and pose-guided association [4] are developed for pose estimation, face recognition, and tracking respectively. In this paper, a robust and distinguishable representation is developed to help the model explicitly learn from the crowded scenes.

3. Our Approach

3.1. Overview

The overall framework is illustrated in Fig. 2. It adopts a simple multi-head design with a backbone and three head networks. Given a single RGB image as input, it outputs a Body Center heatmap, Camera map, and SMPL map, describing the detailed information of the estimated 3D human meshes. In the Body Center heatmap, 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. For simplicity, we combine the Camera map and SMPL map into the Mesh Parameter map. During inference, we sample the 3D body mesh parameter results from the Mesh Parameter map at the 2D body center locations parsed from the Body Center heatmap. Finally, we put the

---

[1]The arXiv version of ROMP, called CenterHMR [46], predates BMP.

sampled parameters into the SMPL model to generate the 3D body meshes.

3.2. Basic Representations

In this section, 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. Here, we set $H = W = 64$.

Body Center heatmap: $C_m \in \mathbb{R}^{1 \times H \times W}$ is a heatmap representing the 2D human body center in the image. Each body center is represented as a Gaussian distribution in the Body Center heatmap. For better representation learning, the Body Center heatmap also integrates the scale information of the 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 Body Center heatmap, k is derived as

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

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.

Mesh Parameter map: $P_m \in \mathbb{R}^{145 \times H \times W}$ 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 its corresponding 3D body mesh parameters. Following [20, 47], we employ a weak-perspective camera model to project K 3D body joints $J = (x_k, y_k, z_k), k = 1 \cdots K$ of the estimated 3D mesh back to the 2D joints $\widehat{J} = (\widehat{x}_k, \widehat{y}_k)$ on the image plane. This facilitates training the model with in-the-wild 2D pose datasets, which helps with robustness and generalization.

Camera map: $A_m \in \mathbb{R}^{3 \times H \times W}$ contains the 3-dim camera parameters (s, t_x, t_y) that describe the 2D scale s and translation $t = (t_x, t_y)$ of the person in the image. The scale s reflects the body size and the depth 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 2D projection $\widehat{J}$ of 3D body joints J can be derived as $\widehat{x}_k = sx_k + t_x, \widehat{y}_k = sy_k + t_y$. The translation parameters allow more accurate position estimates than the Body Center heatmap.

SMPL map: $S_m \in \mathbb{R}^{142 \times H \times W}$ contains the 142-dim SMPL parameters, which describe the 3D pose and shape of the body mesh. SMPL establishes an efficient mapping from the pose θ and shape β parameters to the human 3D body mesh $M \in \mathbb{R}^{6890 \times 3}$. The shape parameter $\beta \in \mathbb{R}^{10}$ is the top-10 PCA coefficients of the SMPL statistical shape space. The pose parameters $\theta \in \mathbb{R}^{6 \times 22}$ contain the 3D rotation of the 22 body joints in a 6D representation [59]. Instead of using the full 24 joints of the original SMPL model, we drop the last two hand joints. The 3D rotation of the first joint denotes the body 3D orientation in the camera coordinate system, while the remainder are the relative 3D ori-
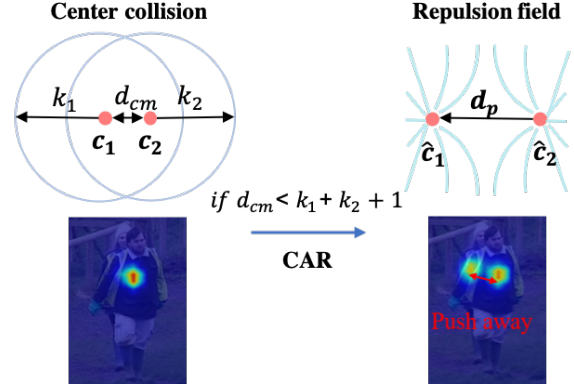

Figure 3. **Collision-Aware Representation.** The body centers of overlapping people are treated as positive charges, which are pushed apart if they are too close in the repulsion field.

entations of each body part with respect to its parent in a kinematic chain. 3D joints J are derived via PM, where $P \in \mathbb{R}^{K \times 6890}$ is a sparse weight matrix that describes the linear mapping from 6890 vertices of the body mesh M to the K body joints.

3.3. CAR: Collision-Aware Representation

The entire framework is based on a concise body-center-guided 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 center-based methods [10, 57] define the center of the bounding box as the target center. This works well for general objects (e.g., a ball or bottle) that lack semantically meaningful keypoints. However, a bounding box center is not a meaningful point on the human body and often falls outside the body area. For stable parameter sampling, we need an explicit body center. Therefore, we calculate each body center from the ground truth 2D pose. Considering that any 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 predict the body location 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 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 distin-

guished. 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 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. Fig. 3 illustrates the principle of CAR. Suppose that $c_1 \in \mathbb{R}^2, c_2 \in \mathbb{R}^2$ are the 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 apart via

$$\hat{c}_1 = c_1 + \gamma d_p, \hat{c}_2 = c_2 - \gamma d_p,$$
$$d_p = \frac{k_1 + k_2 + 1 - d_{cm}}{d_{cm}}(c_1 - c_2), \quad (2)$$

where $d_p \in \mathbb{R}^2$ is the repulsion vector from c_2 to c_1 and γ is an intensity coefficient to adjust the strength. When there are multiple overlapped people, we take Eq. (2) to generate the mutual repulsion vectors d_p^i for the i-th pair of centers. For the center that is affected by N repulsive forces, we calculate the composition of these forces as the numerical summation $\sum_{i=1}^{N} d_p^i$.

During training, we use CAR to push apart close body centers and use them to supervise the Body Center heatmap. In this way, the model is encouraged to estimate the centers that maintain a distinguishable distance. For the Body Center heatmap, it helps the model to effectively locate the occluded person. For the Mesh 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 parse the 3D body meshes from the estimated maps, we need to first parse the 2D body center coordinates $c \in \mathbb{R}^{K \times 2}$ from C_m, where K is the number of the detected people, and then use them to sample P_m for the SMPL parameters. In this section, we introduce the process of the center parsing, matching and sampling.

C_m is a probability map whose local maxima are regarded as the body centers. The local maxima are derived via $Mp(C_m) \wedge C_m$ where Mp is the max pooling operation and $\wedge$ is the logical conjunction operation. Let c be the 2D coordinates of a local maximum with confidence score larger than a threshold t_c. We rank the confidence score at each c and take the top N as the final centers. During inference, we directly sample the parameters from P_m at c.

During training, the estimated 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 by using the center confidence from C_m and the 2D body scale s of the camera parameters from A_m. For people of different s, we regard the one with the larger s as located in the front. For people of similar s, the person with a higher center confidence is considered to be in the front. Please refer to the SuppMat for the details.

3.5. Loss Functions

To supervise ROMP, we develop individual loss functions for different maps. In total, ROMP is supervised by the weighted sum of the body center loss L_c and mesh parameter loss L_p.

Body Center loss. L_c encourages a high confidence value at the body center c of the Body Center heatmap C_m and low confidence elsewhere. To deal with the imbalance between the center location and the non-center locations in C_m, we train the Body Center heatmap based on the focal loss [29]. Given the predicted Body Center heatmap C_m^p and the ground truth C_m^{gt}, L_c is defined as

$$L_c = -\frac{L_{pos} + L_{neg}}{\sum I_{pos}} w_c,$$
$$L_{neg} = log(1 - C_m^p)(C_m^p)^2(1 - C_m^{gt})^4(1 - I_{pos}), \quad (3)$$
$$L_{pos} = log(C_m^p)(1 - C_m^p)^2 I_{pos}, I_{pos} = C_m^{gt} \geq 1,$$

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

Mesh 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 mesh parameter loss is derived as

$$L_p = w_{pose}L_{pose} + w_{shape}L_{shape} + w_{j3d}L_{j3d}$$
$$+ w_{paj3d}L_{paj3d} + w_{pj2d}L_{pj2d} + w_{prior}L_{prior}. \quad (4)$$

L_{pose} is the L_2 loss of the pose parameters 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 J regressed from the body mesh M. L_{paj3d} is the L_2 loss of the 3D joints J after Procrustes alignment. L_{pj2d} is the L_2 loss of the projected 2D joints $\widehat{J}$. L_{prior} is the Mixture Gaussian prior loss of the SMPL parameters adopted in [6, 33] for supervising the plausibility of 3D joint rotation and body shape. Lastly, $w_{(.)}$ denotes the corresponding loss weights.

4. Experiments

4.1. Implementation Details

Network Architecture. For a fair comparison with other approaches, we use ResNet-50 [12] as the default backbone. Since our method is not limited to a specific backbone, we also test HRNet-32 [8] in the experiments. Through the

Method	MPJPE↓	PMPJPE↓	PCK↑	AUC↑	MPJAE↓	PMPJAE↓
OpenPose + SPIN [23]	95.8	66.4	33.3	55.0	23.9	24.4
YOLO + VIBE [22]*	94.7	66.1	33.9	56.6	25.2	20.46
CRMH [15]	105.9	71.8	28.5	51.4	26.4	22.0
BMP [54]*	104.1	63.8	32.1	54.5	-	-
ROMP (ResNet-50)	87.0	62.0	34.4	57.6	21.9	20.1
ROMP (ResNet-50)*	**80.1**	**56.8**	36.4	**60.1**	20.8	19.1
ROMP (HRNet-32)*	82.7	60.5	**36.5**	59.7	**20.5**	**18.9**

Table 1. Comparisons to the state-of-the-art methods on 3DPW following *Protocol 1* (without using any ground truth during inference). * means using extra datasets for training.

Method	MPJPE↓	PMPJPE↓	PVE↓
HMR [20]	130.0	76.7	-
Kanazawa et al. [21]	116.5	72.6	139.3
Arnab et al. [3]	-	72.2	-
GCMR [24]	-	70.2	-
DSD-SATN [47]	-	69.5	-
SPIN [23]	96.9	59.2	116.4
ROMP (ResNet-50)	**91.3**	**54.9**	**108.3**
I2L-MeshNet [37]*	93.2	58.6	-
EFT [19]*	-	54.2	-
VIBE [22]*	93.5	56.5	113.4
ROMP (ResNet-50)*	89.3	53.5	105.6
ROMP (HRNet-32)*	**85.5**	**53.3**	**103.1**

Table 2. Comparisons to the state-of-the-art methods on 3DPW following VIBE [22], using *Protocol 2* (on the test set only). * means using extra datasets (compared with SPIN) for training.

backbone, a feature vector $f_b \in \mathbb{R}^{32 \times H_b \times W_b}$ is extracted from a single RGB image. Also, we adopt the Coord-Conv [31] to enhance the spatial information. Therefore, the backbone feature $f \in \mathbb{R}^{34 \times H_b \times W_b}$ is the combination of a coordinate index map $ci \in \mathbb{R}^{2 \times H_b \times W_b}$ and f_b. Next, from f, three head networks are developed to estimate the Body Center, Camera, and SMPL maps. More details of the architecture are in the SuppMat.

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 number of detections, $N = 64$, which is set manually. The loss weights are set to $w_c = 160, w_{j3d} = 360, w_{paj3d} = 400, w_{pj2d} = 420, w_{pose} = 80, w_{shape} = 1$, and $w_{prior} = 1.6$ to ensure that the weighted loss items are of the same magnitude. The threshold t_c of the Body Center heatmap is 0.2. The repulsion coefficient γ of CAR is 0.2.

Training Datasets. For a fair comparison with previous methods [15, 20, 23, 47], the basic training datasets we used in the experiments include two 3D pose datasets (Human3.6M [14] and MPI-INF-3DHP [34]), one pseudo-label 3D dataset (UP [26]) and four in-the-wild 2D pose datasets (MS COCO [30], MPII [2], LSP [16, 17] and AICH [50]).

We also use the pseudo 3D annotations from [23]. To further explore the upper limit of performance, we also use additional training datasets, including two 3D pose datasets (MuCo-3DHP [34] and OH [55]), the pseudo 3D labels of 2D pose datasets provided by [19], and two 2D pose datasets (PoseTrack [1] and Crowdpose [28]), to train an advanced model.

Evaluation Benchmarks. 3DPW [49] is employed as the main benchmark for evaluating 3D mesh/joint error since it contains in-the-wild multi-person videos with abundant 2D/3D annotations. Specially, we divide 3DPW into 3 subsets, including 3DPW-PC for person-person occlusion, 3DPW-OC for object occlusion, and 3DPW-NC for the non-occluded/truncated cases, to evaluate the performance in different scenarios. Additionally, we also evaluate on a indoor multi-person 3D pose benchmark, CMU Panoptic [18]. Furthermore, we evaluate the stability under occlusion on Crowdpose [28], a crowded-person in-the-wild 2D pose benchmark.

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 (PMPJPE), percentage of correct keypoints (PCK), and area under the PCK-threshold curve (AUC). We also adopt mean per joint angle error (MPJAE), and Procrustes-aligned MPJAE (PA-MPJAE) to evaluate the 3D joint rotation accuracy. Also, to evaluate the pose accuracy in crowded scenes, we calculate the average precision ($\text{AP}^{0.5}$) between the 2D-projection $\hat{J}$ and the ground truth 2D poses on Crowdpose.

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

3DPW. We adopt three evaluation protocols, which reveal different properties. To validate the performance in actual scenes, we follow *Protocol 1* from the 3DPW Challenge to evaluate on the entire 3DPW dataset without using any ground truth, especially the bounding box. With the whole image as input, we equip each single-person method [22, 23] with a human detector (OpenPose [7] or YOLO [41]). For a fair comparison, ROMP uses the same backbone (ResNet-50) and training datasets as the competing method [23]. We obtain the results of OpenPose + SPIN

Method	MPJPE↓	PMPJPE↓	PVE↓
EFT [19]	-	52.2	-
VIBE [22]*	82.9	51.9	99.1
ROMP (ResNet-50)	84.2	51.9	100.4
ROMP(ResNet-50)*	79.7	49.7	94.7
ROMP (HRNet-32)	78.8	48.3	94.3
ROMP (HRNet-32)*	**76.7**	**47.3**	**93.4**

Table 3. Comparisons to the state-of-the-art methods on 3DPW following *Protocol 3* (fine-tuned on the training set). * means using extra datasets (compared with EFT) for training.

Method	Haggling	Mafia	Ultim.	Pizza	Mean
Zanfir et. al. [53]	141.4	152.3	145.0	162.5	150.3
MSC [52]	140.0	165.9	150.7	156.0	153.4
CRMH [15]	129.6	133.5	153.0	156.7	143.2
ROMP (ResNet-50)	**111.8**	**129.0**	148.5	149.1	**134.6**
BMP [54]*	120.4	132.7	140.9	147.5	135.4
ROMP (ResNet-50)*	**107.8**	125.3	**135.4**	141.8	**127.6**
ROMP (HRNet-32)*	110.8	**122.8**	141.6	**137.6**	128.2

Table 4. Comparisons to the state-of-the-art methods on CMU Panoptic [18] benchmark. The evaluation metric is MPJPE after centering the root joint. All methods are directly evaluated without any fine-tuning. * means using extra datasets for training.

Method	3DPW-PC	3DPW-NC	3DPW-OC
CRMH [15]	103.5	68.7	78.9
VIBE [22]	103.9	57.3	**65.9**
ROMP w/o CAR	79.7	56.7	67.0
- w/ CAR ($\gamma = 0.1$)	77.6	**55.6**	66.6
- w/ CAR ($\gamma = 0.2$)	**75.8**	57.1	67.1
- w/ CAR ($\gamma = 0.3$)	77.0	56.4	66.5

Table 5. Comparisons to the state-of-the-art methods on the person-occluded (3DPW-PC), object-occluded (3DPW-OC) and non-occluded/truncated (3DPW-NC) subsets of 3DPW. We also ablate CAR and vary the repulsion coeff. γ. The evaluation metric is PMPJPE.

Split	CRMH [15]	ROMP	ROMP+CAR
Test	33.9	54.1	**59.7**
Validation	32.9	55.6	**58.6**

Table 6. Comparisons to the state-of-the-art methods on the Crowdpose [28] benchmark. The evaluation metric is $\text{AP}^{0.5}$.

Method	VIBE [22]	CRMH [15]	ROMP	ROMP
FPS ↑	10.9	14.1	20.8	**30.9**
Backbone	ResNet-50	ResNet-50	HRNet-32	ResNet-50

Table 7. Run-time comparisons on a 1070Ti GPU.
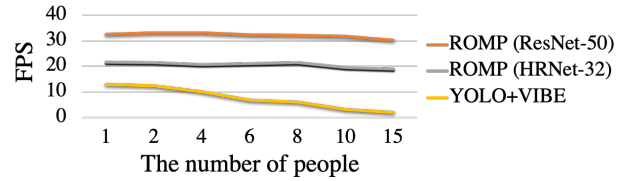

Figure 4. The FPS variations of ROMP and YOLO+VIBE [22] when processing images with different number of people.

from [13]. The results of YOLO + VIBE are obtained using the officially released code, which already contains the YOLO part for human detection. The results of BMP, which adopts a multi-scale grid-level representation, are obtained from [54]. In Tab. 1, ROMP significantly outperforms all these methods, particularly in MPJPE, PMPJPE, and MP-JAE. These results validate that learning a robust pixel-level representation with a holistic view is helpful for improving the robustness and generalization in actual scenes. Training with extra datasets (∗) shows that the accuracy of ROMP can be further improved.

As a sanity check, we also compare ROMP with the single-person approaches in the evaluation protocols that allow them to use the cropped single-person image as input, while ROMP still takes the whole image as input. Follow-ing VIBE [22], *Protocol 2* uses the 3DPW test set for evaluation without fine-tuning on the training set, while *Protocol 3* fine-tunes the model on the 3DPW training set and uses the test set for evaluation. In Tab. 2, ROMP outperforms these multi-stage approaches on *Protocol 2*, further demonstrating the advantage of our one-stage design. In Tab. 3, ROMP achieves comparable results with the state-of-the-art methods. If we use HRNet-32 as the backbone, the accuracy improves significantly after fine-tuning.

CMU Panoptic. Following the evaluation protocol of CRMH [15], we evaluate ROMP on the multi-person benchmark, CMU Panoptic, without any fine-tuning. For a fair comparison, we use the same backbone and similar training set as CRMH. As shown in Tab. 4 , ROMP outperforms the existing multi-stage methods [15, 52, 53] in all activities by a large margin. These results further demonstrate that learning pixel-level representation with a holistic view improves the performance on multi-person scenes.

Occlusion benchmarks. To validate the stability under occlusion, we evaluate ROMP on multiple occlusion benchmarks. Firstly, on the person-occluded **3DPW-PC** and **Crowdpose** [28], results in Tab. 5 and 6 show that ROMP significantly outperforms previous state-of-the-art methods [15, 22]. Additionally, in Fig. 5, some qualitative comparisons to CRMH also demonstrate ROMP's robustness to person-person occlusion. These results suggest that the pixel-level representation is important for improving the performance under person-person occlusion. Finally, on the object-occluded 3DPW-OC, ROMP also achieves promising performance. These results demonstrate that the fine-grained pixel-level representation is beneficial for dealing with various occlusion cases.
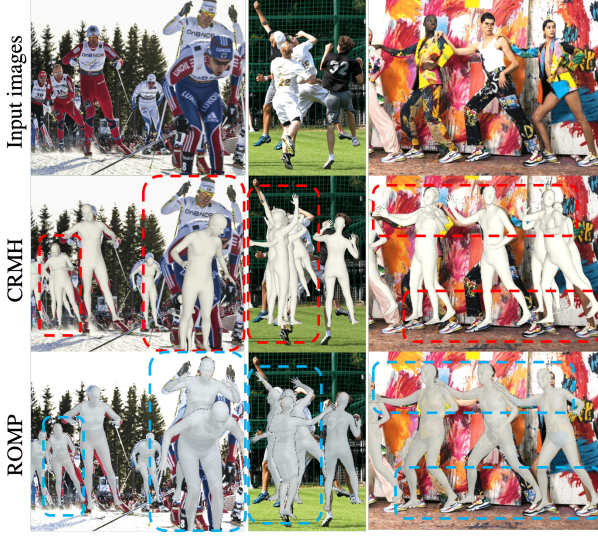
Figure 5. Qualitative comparisons to CRMH [15] on the Crowdpose and the internet images.

Figure 6. Qualitative ablation study of the CAR on the Crowdpose.

Runtime comparisons. We compare ROMP with the state-of-the-art methods in processing videos captured by a web camera. All runtime comparisons are performed on a desktop with a GTX 1070Ti GPU, Intel i7-8700K CPU, and 8 GB RAM. As shown in Tab. 7, ROMP achieves real-time performance, significantly faster than the competing methods. Additionally, as shown in Fig. 4, compared with the multi-stage methods [15, 22], ROMP's processing time is roughly constant regardless of the number of people.

4.3. Ablation Study of the CAR

As shown in Tab. 5 and 6, CAR improve the PMPJPE metric on the 3DPW-PC and the Crowdpose datasets by 4.8% and 10.3%, respectively. Additionally, Fig. 6 shows, qualitatively, the impact of ablating CAR. Adding CAR improves performance in crowded scenes, which demonstrates that CAR effectively alleviates the center collision problem.

Intensity coefficient γ of the CAR. To set γ, we conduct an ablation study on 3DPW-PC. In Tab. 5, setting $\gamma = 0.2$ performs much better on the crowded scenes (3DPW-PC) and its performance on the normal cases (3DPW-NC/3DPW-OC) is comparable to the best. For general in-the-wild cases of Crowdpose, setting $\gamma = 0.2$ improves the performance by 10% over $\gamma = 0$ in Tab. 6. Therefore, we suggest training the model with $\gamma = 0.2$ for all cases. The reason performance degrades in the normal cases is probably that pushing apart the body centers affects the consistency of the body-center-guided representation.

4.4. Discussion

To understand the source of our performance gains, we conduct an ablation study on different subsets of 3DPW subsets. Results in Tab. 5 show that our main gains come
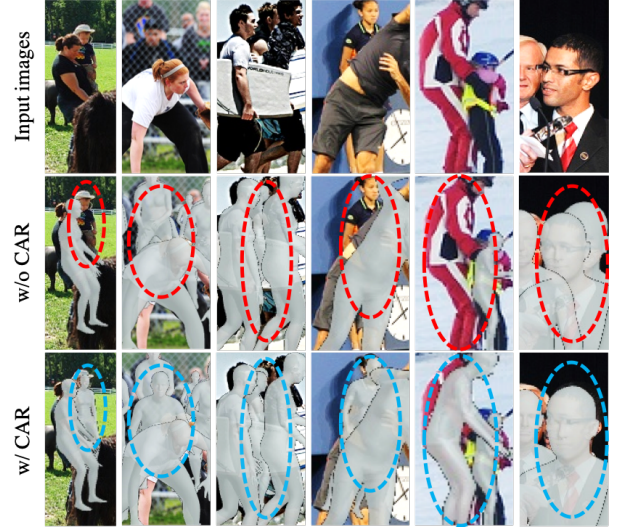
from the person-occluded and the non-occluded/truncated cases. It demonstrates the effectiveness of the proposed pixel-level representation in improving the disambiguation in the crowded scenes. Our experiments suggest that the difference between ROMP and the state-of-the-arts [15, 22, 23] is the method of representation learning. ROMP learns the pixel-level representation from a holistic view, while the multi-stage methods learn a bounding-box-level representation in a local view. Our one-stage framework enables ROMP to learn more discriminative features that are robust to rich disturbances outside the bounding box, helping generalization.

5. Conclusion

We introduce a novel one-stage network, ROMP, for monocular multi-person 3D mesh regression. For pixel-level estimation, we propose an explicit body-center-guided representation and further develop it as a collision-aware version, CAR, enabling robust prediction under person-person occlusion. ROMP is the first open-source one-stage method that achieves state-of-the-art performance on multiple benchmarks as well as real-time inference speed. ROMP can serve as a simple yet effective foundation for related multi-person 3D tasks, such as depth estimation, tracking, and interaction modeling.

References

[1] Mykhaylo Andriluka, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele. PoseTrack: A benchmark for human pose estimation and tracking. In *CVPR*, 2018. 6

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[3] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3D human pose estimation in the wild. In *CVPR*, 2019. 6

[4] Qian Bao, Wu Liu, Yuhao Cheng, Boyan Zhou, and Tao Mei. Pose-guided tracking-by-detection: Robust multi-person pose tracking. *IEEE Transactions on Multimedia*, 23:161–175, 2020. 3

[5] 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

[6] 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. 2, 5

[7] 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, 6

[8] 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, 5

[9] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *ECCV*, 2020. 2

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

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

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

[13] 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

[14] 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

[15] 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

[16] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 6

[17] Sam Johnson and Mark Everingham. Learning effective human pose estimation from inaccurate annotation. In *CVPR*, 2011. 6

[18] 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

[19] 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. 6, 7

[20] 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, 4, 6

[21] Angjoo Kanazawa, Jason Zhang, Panna Felsen, and Jitendra Malik. Learning 3D human dynamics from video. In *CVPR*, 2019. 1, 2, 6

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

[23] 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, 6, 8

[24] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 1, 6

[25] Jogendra Nath Kundu, Mugalodi Rakesh, Varun Jampani, Rahul Mysore Venkatesh, and R Venkatesh Babu. Appearance consensus driven self-supervised human mesh recovery. In *ECCV*, 2020. 2

[26] 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 *CVPR*, 2017. 6

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

[28] 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

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

[30] 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. 6

[31] 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

[32] Wu Liu, Qian Bao, Yu Sun, and Tao Mei. Recent advances in monocular 2d and 3d human pose estimation: A deep learning perspective. *arXiv*, 2021. 2

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *TOG*, 2015. 2, 3, 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 *3DV*, 2017. 6

[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 *3DV*, 2018. 2

[36] 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. 2

[37] 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. 6

[38] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NeurIPS*, 2017. 3

[39] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. TexturePose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 1, 2

[40] 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

[41] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv*, 2018. 6

[42] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 3

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

[44] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015. 3

[45] Konstantin Sofiiuk, Olga Barinova, and Anton Konushin. Adaptis: Adaptive instance selection network. In *ICCV*, 2019. 3

[46] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Mei Tao. Centerhmr: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image. In *arxiv*, 2021. 3

[47] 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. 1, 2, 4, 6

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

[49] 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

[50] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, et al. Ai challenger: A large-scale dataset for going deeper in image understanding. *arXiv*, 2017. 6

[51] 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

[52] 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

[53] 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

[54] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, 2021. 3, 6, 7

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

[56] 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 *ECCV*, 2020. 3

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

[58] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *CVPR*, 2019. 3

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

[60] 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