Single-Stage is Enough: Multi-Person Absolute 3D Pose Estimation

Lei Jin[1], Chenyang Xu[1], Xiaojuan Wang[1†], Yabo Xiao[1], Yandong Guo[2], Xuecheng Nie[3], Jian Zhao[4†]
[1]Beijing University of Posts and Telecommunications
[2]OPPO Research Institute
[3]National University of Singapore
[4]Institute of North Electronic Equipment
{jinlei,xuchenyang,wj271,xiaoyabo}@bupt.edu.cn, guoyandong@oppo.com, niexuecheng@u.nus.edu,
zhaojian90@u.nus.edu [*]

Abstract

The existing multi-person absolute 3D pose estimation methods are mainly based on two-stage paradigm, i.e., top-down or bottom-up, leading to redundant pipelines with high computation cost. We argue that it is more desirable to simplify such two-stage paradigm to a single-stage one to promote both efficiency and performance. To this end, we present an efficient single-stage solution, Decoupled Regression Model (DRM), with three distinct novelties. First, DRM introduces a new decoupled representation for 3D pose, which expresses the 2D pose in image plane and depth information of each 3D human instance via 2D center point (center of visible keypoints) and root point (denoted as pelvis), respectively. Second, to learn better feature representation for the human depth regression, DRM introduces a 2D Pose-guided Depth Query Module (PDQM) to extract the features in 2D pose regression branch, enabling the depth regression branch to perceive the scale information of instances. Third, DRM leverages a Decoupled Absolute Pose Loss (DAPL) to facilitate the absolute root depth and root-relative depth estimation, thus improving the accuracy of absolute 3D pose. Comprehensive experiments on challenging benchmarks including MuPoTS-3D and Panoptic clearly verify the superiority of our framework, which outperforms the state-of-the-art bottom-up absolute 3D pose estimation methods.

1. Introduction

Estimating 3D human pose from a monocular RGB camera is a significant task in computer vision and artificial intelligence, due to its foundation in many higher-level applications, *e.g.*, robotics [41], action recognition [8, 15], animation [36, 37], human-object interaction detection [6, 12, 38], virtual fitting [11], *etc*. With the recent notable progress in single-person based 3D pose estimation [3, 7, 23, 25, 33], a more realistic and challenging problem setting has attracted increasing attention, *i.e.*, to estimate 3D human pose for multiple persons from a single image.

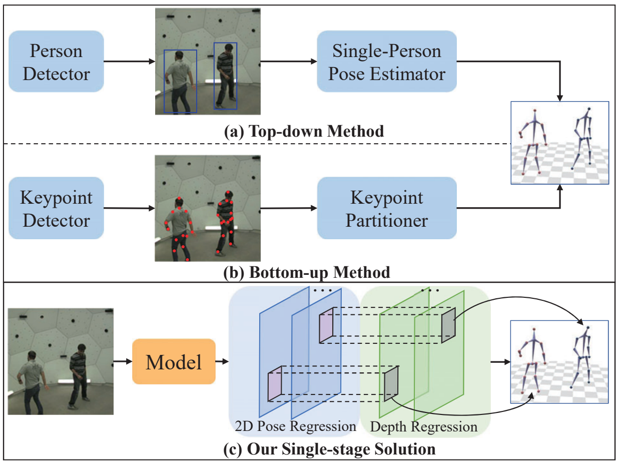

Figure 1. Comparison between our single-stage solution and existing top-down and bottom-up methods for the multi-person 3D pose estimation.

In general, existing multi-person 3D pose estimation paradigms can be classified as top-down and bottom-up methods, as illustrated in Fig. 1 (a), (b), respectively. Top-down approaches [1, 13, 20, 30] use a human detector to obtain the bounding box of each person, and then perform the single-person pose estimation, while bottom-up approaches [19, 39] estimate the poses of all persons simultaneously, and then combine the keypoints belonging to the same person. The former category estimates pose for each person separately, hence the total computation cost grows linearly with the number of people in the image; the latter category requires grouping the keypoints into corresponding persons, leading to redundant computational complexity.

Despite the recent popularity and promising perfor-

[*]Corresponding authors: Xiaojuan Wang and Jian Zhao.

mance of the single-stage methods for 2D pose estimation [22,31,32,40], the single-stage pipeline for multi-person 3D pose estimation is barely explored, as it remains unclear how to effectively combine the end-to-end 2D pose regression with person depth estimation. In this paper, we propose a single-stage pipeline, termed as Decoupled Regression Model (DRM). DRM introduces a new decoupled formulation, which represents the 2D pose and depth information of each 3D human instance via 2D center point (center of visible keypoints) and root point (denoted as pelvis). Specifically, we perform 2D keypoint regression from 2D center point and keypoint depth estimation from root point via two parallel branches, thus effectively unifying the 2D pose regression with person depth estimation to jointly perform 3D pose regression.

Since measuring depth from a single image is ambiguous, estimating absolute 3D pose naturally suffers from the inaccurate human depth estimation. Considering that the features used for the absolute depth prediction need to adequately perceive the high-level features, *e.g.*, human scale, relative location, *etc*. From the perspective camera model, the human scale and location can partly describe depth information. To learn better feature representation for distinguishing instances at different depth, DRM introduces a plug-in 2D Pose-guided Depth Query Module (PDQM) to extract the features in the 2D pose regression branch, which is experimentally proved to be beneficial to absolute depth prediction. Specifically, we design a warp operation to query features from the positions of predicted 2D poses, and then concatenate these features to that of the depth to enhance the depth prediction branch. Also, in order to further improve the accuracy of estimation for both the root absolute depth and the root-relative depth, we propose a Decoupled Absolute Pose Loss (DAPL) to supervise the human absolute 3D pose in the camera coordinate system. It is proved that DAPL can further advance the improvements brought by PDQM. Comprehensive experiments on the challenging 3D pose benchmarks MuPoTS-3D [18] and Panoptic [9] evidently demonstrate the superior efficacy of the proposed DRM.

Our main contributions are summarized as follows.

- We propose the first single-stage solution Decoupled Regression Model (DRM) for multi-person absolute 3D pose estimation, which decomposes the problem-to-solve into 2D pose regression and depth regression via decoupled representation.
- DRM introduces a plug-in 2D Pose-guided Depth Query Module (PDQM) to inject the features of the 2D pose regression branch to the depth regression branch through a position query operation, which helps our model adaptively perceive the scale information of instances.
- DRM also introduces a Decoupled Absolute Pose Loss (DAPL) to focus on the absolute depth prediction, which serves as a supplement to the PDQM.
- DRM achieves comparable performance with the most top-down methods and significantly outperforms the state-of-the-art bottom-up method [39] by 4.6 PCK$_{rel}$ and 2.3 PCK$_{abs}$ on the MuPoTS-3D [18] and 4.9 MPJPE on Panoptic [9] benchmarks, respectively.

2. Related Work

Single-Person 3D Pose Estimation There are two lines to solve the problem of single-person 3D pose estimation with monocular RGB images: single-stage [10, 24, 27, 28] and two-stage [16, 21, 33] approaches. The single-stage approaches directly locate 3D human keypoints from the input image. For example, Pavlakos *et al.* [24] propose a coarse-to-fine approach to estimate a 3D heatmap for pose estimation. Kanazawa *et al.* [10] propose end-to-end adversarial learning of 3D pose and body mesh by minimizing the reprojection loss. Sun *et al.* [28] formulate an integral operation as soft-argmax to obtain 3D pose coordinates in a differentiable manner. Differently, the two-stage approaches first predict 2D poses by utilizing an off-the-shelf accurate 2D pose estimator, and then lift them to the 3D space. For instance, Martinez *et al.* [16] propose a simple baseline to regress 3D pose from 2D coordinates directly. Moreno-Noguer [21] obtains more precise pose estimation by the distance matrix representation. Yang *et al.* [33] utilize a multi-source discriminator to generate anthropometrically valid poses.

Multi-person 3D Pose Estimation For multi-person 3D pose estimation with monocular RGB images, similar categories as multi-person 2D pose estimation are noted: top-down [1, 2, 13, 20, 30] and bottom-up [19, 39] approaches. The top-down approaches first perform human detection to detect each individual person, then for each detected person instance, absolute root (pelvis of the human) depth and 3D root-relative pose are estimated by 3D pose estimation models. For instance, Moon *et al.* [20] introduce a camera distance-aware approach that a cropped human image is fed into their designed RootNet to estimate the body's root depth, then the root-relative 3D pose is estimated by their proposed PoseNet. Benzine *et al.* [1] propose a single-shot approach and introduce a low-resolution anchor-based representation learning scheme to avoid the occlusion problem. Li *et al.* [30] adopt a hierarchical multi-person ordinal relations method to leverage body level semantic and global consistency for encoding the interaction information hierarchically. Lin *et al.* [13] formulate human depth regression as a bin index estimation problem for multi-person localization in the camera coordinate system. In contrast, bottom-up approaches first predict all body keypoint locations and depth maps, then associate body parts to each person according to the root depth and root-relative depth. For example, Mehta *et al.* [19] infer intermediate 3D pose of visible body keypoints regardless of the accuracy, then the completed 3D pose is reconstructed by inferring occluded keypoints using learned pose priors and global context. The final 3D pose is refined by applying temporal coherence and fitting the kinematic skeletal model. Zhen *et al.* [39] leverage a depth-aware part association algorithm to assign keypoints to individuals by reasoning about inter-person occlusion and bone-length constraints.

Monocular Depth Estimation For depth estimation in multi-person absolute 3D pose estimation, most methods [20, 39] use a sparse depth map to supervise the depth value in 2D position of the root (set in pelvis) point. Differently, Zhang *et al.* [35] discretize depth into several levels to represent the depths of instances and use ordinal depth relations among instances to supervise the depth ordering. We argue that human depth estimation should perceive global features related to the scale of instance. Hence, we propose to inject the features of the 2D pose regression branch to the depth regression branch.

The above elaboration states that the two-stage methods for multi-person 3D pose estimation have their disadvantages, respectively. The top-down methods highly depend on the performance of the human detector and barely have good strategy to solve the problem of occlusion, while the bottom-up methods rely on the grouping algorithms after obtaining the complex intermediate representations to recover poses of all people. Whereas our single-stage DRM manifests comparable accuracy to top-down methods and a more compact pipeline to bottom-up methods.

3. Decoupled Regression Model

In this paper, we aim at proposing a single-stage method which is capable of achieving comparable performance with that of the two-stage's for multi-person 3D pose estimation in a more efficient and compact pipeline. The proposed Decoupled Regression Model (DRM) has a better tradeoff between performance and computational complexity without any bells and whistles.

3.1. Decoupled Representation for 3D Pose

Given an image I, the multi-person absolute 3D pose estimation is to locate human keypoints of all the person instances $\mathcal{P} = \left\{P_m^{abs}\right\}_{m=1}^{N}$ in I, where N denotes the number of persons in I. Assume that there are J keypoints in a single 3D pose skeleton. The $m\text{-}th$ absolute 3D pose can be formulated as: $P_m^{abs} = \left\{\left(X_{m,j}^{abs}, Y_{m,j}^{abs}, Z_{m,j}^{abs}\right)^T\right\}_{j=1}^{J}$, where $\left(X_{m,j}^{abs}, Y_{m,j}^{abs}, Z_{m,j}^{abs}\right)^T$ is the $j\text{-}th$ keypoint position of the $m\text{-}th$ absolute pose in the camera-centered coordinate system, as shown in Fig. 2 (c).

2D poses $\left\{p_m\right\}_{m=1}^{N}$, root-relative depth $\left\{\varDelta Z_m\right\}_{m=1}^{N}$, and absolute depth of the root point $\left\{Z_{m,r}\right\}_{m=1}^{N}$ are needed to estimate the absolute 3D poses in the DRM. The $m\text{-}th$ 2D pose p_m and root-relative depth $\varDelta Z_m$ are formulated as:

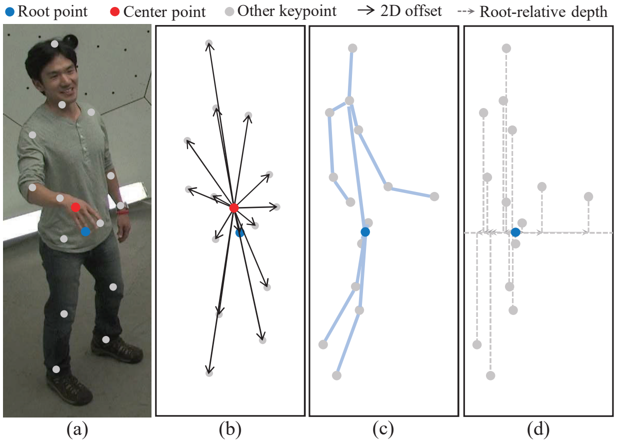

Figure 2. Visualization and explanation of the pose representation. (a) Pose representation overlaid on an image containing a person instance. (b) 2D pose representation based on the center point. (c)3D pose for the instance in the right view. (d) Relative depth representation based on the root point. In our framework, the root point and the center point are different.

$$p_m = \left\{\left(x_{m,j}, y_{m,j}\right)^T\right\}_{j=1}^{J}, \quad (1)$$

$$\varDelta Z_m = \left\{Z_{m,r} - Z_{m,j}\right\}_{j=1}^{J}, \quad (2)$$

where $\left(x_{m,j}, y_{m,j}\right)^T$ is the $j\text{-}th$ keypoint position of the $m\text{-}th$ 2D pose in the pixel coordinates, and $Z_{m,j}$ is the $j\text{-}th$ keypoint absolute depth of the $m\text{-}th$ instance.

Hence, we decompose multi-person 3D pose estimation into two simultaneous regression-based tasks, *i.e.*, 2D pose regression and depth regression. Furthermore, we adopt the center point and root point as the regression clue for 2D pose regression and depth regression, respectively.

2D Pose Regression We use a center map $\mathcal{C}$ and n offset maps $\mathcal{O}$ to locate instances in the given image I, as shown in Fig. 2 (b). The center map is modeled as a Gaussian-based heatmap, whose values represent the confidence of the center position. We denote the groundtruth center map with $\mathcal{C}^*$. We set the instance center point at the average coordinate of all visible keypoints of the instance, and the center is the regression clue for the 2D pose regression branch in DRM. For the position (x, y) in I, $C^*(x, y) = \exp\left(-\|(x, y) - (x^c, y^c)\|^2/\sigma^2\right)$, where (x^c, y^c) is the position of an instance center, and σ is the Gassian variance. Each of the offset maps $\mathcal{O}$ predicts a $2n$-dimension offset vector from the center pixel q for n keypoints at each center pixel q of all instances. The groundtruth offset maps $\mathcal{O}^*$ for each image are constructed from all the 2D poses $\{p_1, p_2, \cdots, p_n\}$ in the image. We compute the center position $\overline{p_i} = \frac{1}{n}\sum_{k=1}^{n} p_{ik}$. The candidate area is around the center position and its radius is

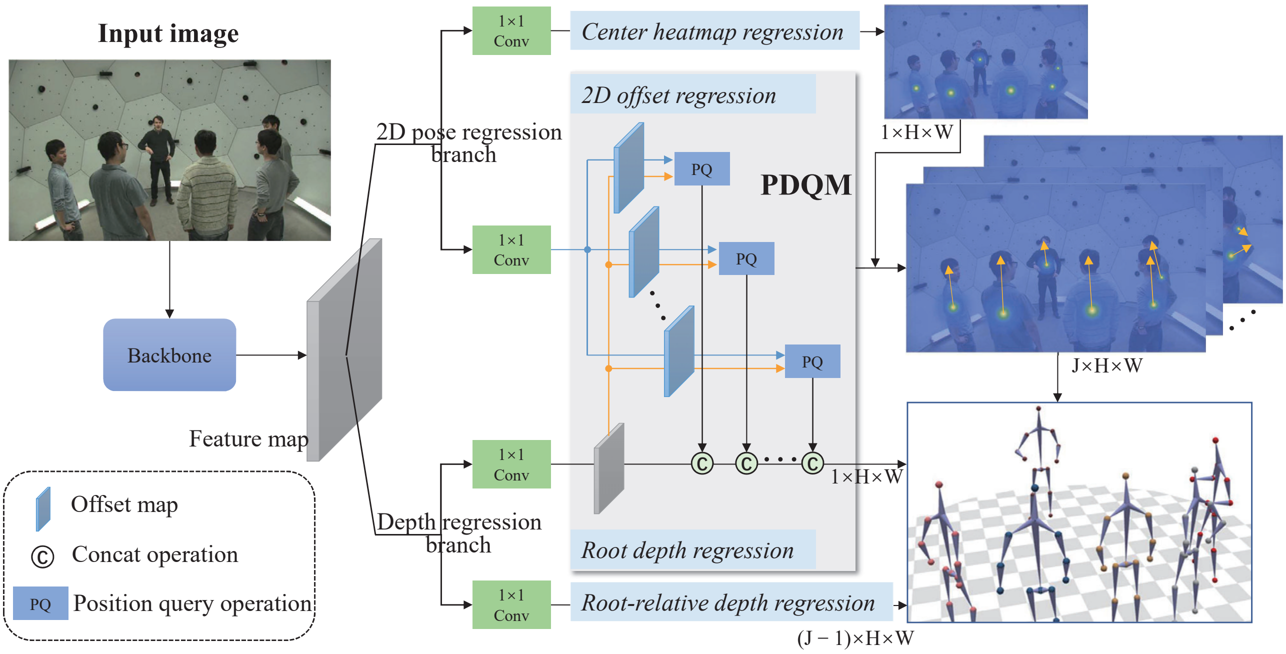

Figure 3. Overview of the proposed DRM for multi-person absolute 3D pose estimation. Given an input monocular image, our single-stage network is divided into four flows, which output a center map, offset maps, a root depth map and root-relative depth maps, respectively. Notablely, with the proposed PDQM, the features for 2D offset regression are shared in the root depth regression via the concatenate operations. Absolute 3D poses of all people can be reconstructed through all these regression maps.

set to 3 according to previous method [40]. The pixels in the candidate area represent the offset to the keypoints $\{\overline{p_i} - p_{i1}, \overline{p_i} - p_{i2}, \cdots, \overline{p_i} - p_{in}\}$.

Depth Regression Instead of predicting the absolute depth values for all keypoints, we only regress the absolute depth of root point and the root-relative depth of other keypoints, as shown in Fig. 2 (d). Such representation makes our depth regression retain relative information for body keypoints and improves the overall training stability. The groundtruth absolute depth is represented by a dense depth map $\mathcal{Z}^*$ at the root pixel r, whose value indicates the groundtruth depth of root point. Similarly, the groundtruth root-relative depth is represented by $(n-1)$-dimension dense depth maps $\mathit{\Delta}\mathcal{Z}^*$ to encode the differences of depth between other keypoints and root point at each root pixel r of all instances. We set the root point at the pelvis.

In this way, the 2D pose and the depth are decoupled, which prevents them from influencing each other. To achieve the final result, we combine the 2D pose and the depth prediction through the root point.

Relation to Previous Representations In existing researches [20, 39], decoupling 3D pose estimation as 2D pose estimation and depth prediction has also been explored. Unlike their decoupled form only in task level, we further decouple the clue keypoints, using the center point and root point for 2D pose regression and corresponding depth regression, respectively, the efficacy of the decoupled representation is experimentally analyzed in Sec. 4.2. Benefiting from the decoupled representation in the clue keypoints, the two regression branches both achieve better performance compared to previous methods.

3.2. Framework Architecture

The framework overview of the proposed single-stage DRM is illustrated in Fig. 3. First, an input image I is sent into the backbone to produce a feature map $\mathcal{X}$. Then $\mathcal{X}$ is transformed into four intermediate supervision flows. One flow is to regress the center map, which contains 1 channel. The other flow is for the offset maps with $2n$ channels, consisting of x-axis and y-axis offsets for n keypoints. The rest two flows are set to regress the depth maps, *i.e.*, 1 channel for the absolute root depth and $n-1$ channels for the root-relative depth, consisting of $n-1$ keypoints except the root point. The disentangled form [40] is adopted to regress the offset maps and the root-relative depth maps.

2D Pose-guided Depth Query Module Depth estimation from a single view suffers from inherent ambiguity. Directly estimating the absolute depth via the feature representations learnt from the whole image is non-trivial, since it is focused on the root area without perceiving global features related to the scale of instances. In fact, the absolute depth of people can be partially expressed by the scale of people. Hence we consider the 2D pose can help advance the absolute depth estimation. To predict the depth of root, we can utilize the features located at other keypoints. Motivated by this, we propose a 2D Pose-guided Depth Query Module (PDQM).

In the flow of offset map regression, we divide the feature maps $\mathcal{X}$ output from the backbone into n feature maps, $\{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_n\}$, and estimate the offset map $\{O_i\}, i = 1, 2, \cdots, n$ for each keypoint from the corresponding feature map:

$$O_i = \mathcal{F}_i(\mathcal{X}_i), i = 1, 2, \cdots, n, \tag{3}$$

where $\mathcal{F}_i(\cdot)$ is the $i\text{-}th$ regressor of the $i\text{-}th$ keypoint, and O_i is the offset map for the $i\text{-}th$ keypoint. The n regressors have the same structure, and they predict their corresponding keypoint offset map independently.

In the flow of root depth regression, we aim to enrich the feature of root by extracting the features around each keypoint. Through the regressed 2D offset maps, we leverage the position query (PQ) operation to extract the $(n-1)$ 64-channel features in the area of each keypoint and concatenate them to the feature map of the root depth regression:

$$\mathcal{Z} = \mathcal{F}_Z\left(cat\left\{\mathcal{X}, W\left(O_1'\right), W\left(O_2'\right)\cdots W\left(O_{n-1}'\right)\right\}\right), \tag{4}$$

where $\mathcal{F}_Z(\cdot)$ is the regressor of root absolute depth, following the same structure as the regressor of offset maps, $cat\{\cdot\}$ is the concatenate operation on channel dimension, $W(O_i')$ is the position query operation on the $i\text{-}th$ keypoint, which is a warp operation to fetch the feature in corresponding position, O_i' is the $i\text{-}th$ offset map from root point to other keypoints, *i.e.*, $O_i' = O_i + O_e$, where O_e is an extra offset map to predict the displacement between root point and center point. We use $\mathcal{Z}$ for the root point depth prediction.

3.3. Training and Inference

We use different losses in each flow. For the 2D pose prediction, the center map loss and the offset map loss are adopted, while for the depth prediction, a novel Decoupled Absolute Pose Loss (DAPL) is designed as the supplement to dense depth map loss.

Center Map Loss The center confidence map is constructed by modeling the center position as Gaussian peaks and the loss function of center map is formulated as the weighted distances between the predicted heat values and the groundtruth heat values:

$$\mathcal{L}_{\mathrm{c}} = \|C - C^*\|_2^2, \tag{5}$$

where $\|\cdot\|_2$ is the entry-wise 2-norm, C and C^* are predicted and target center maps, respectively.

Offset Map Loss The offset maps estimate the candidate poses at each center pixel, by predicting a $2n$-dimension offset vector from the center. We use the smooth ℓ_1 loss for the dense offset maps:

$$\mathcal{L}_{\mathrm{o}} = \sum_{i\in\mathcal{S}} \frac{1}{B_i} smooth_{\ell_1}(o_i - o_i^*), \tag{6}$$

where $\mathcal{S}$ is the set of the positions with groundtruth poses, $B_i = \sqrt{H_i^2 + W_i^2}$ is the size of the corresponding instance, H_i and W_i are the height and the width of the instance box, o_i and o_i^* are the predicted and groundtruth offset for the position i, respectively.

Depth Loss There are two output flows of DRM for depth regression, including absolute depth for the root point and root-relative depth for other keypoints. We use the smooth ℓ_1 loss to formulate the pixel-wise depth loss:

$$\mathcal{L}_{\mathrm{rz}} = \sum_{i\in\mathcal{S}} smooth_{\ell_1}(z_i - z_i^*), \tag{7}$$

$$\mathcal{L}_{\Delta\mathrm{z}} = \sum_{i\in\mathcal{S}} smooth_{\ell_1}(\Delta z_i - \Delta z_i^*), \tag{8}$$

where $\mathcal{S}$ is the set of the positions with groundtruth poses, z_i, a column of $\mathcal{Z}$, is the 1-dimension estimated root depth vector for the position i, and z_i^*, a column of $\mathcal{Z}^*$, is the 1-dimension groundtruth root depth vector for the position i, Δz_i, a column of $\Delta\mathcal{Z}$, is the $(n-1)$-dimension predicted root-relative depth vector for the position i, and Δz_i^*, a column of $\Delta\mathcal{Z}^*$, is the $(n-1)$-dimension groundtruth root-relative depth vector for the position i.

Decoupled Absolute Pose Loss Thanks to the disentangled regression method [40], the 2D poses predicted by our network are accurate enough in most cases while the performance of the estimated absolute depth is poor. To further optimize the absolute pose, we design a Decoupled Absolute Pose Loss (DAPL), which focuses on the absolute depth and the relative depth. Considering that the root-relative depth is local and independently estimated, hence, the estimated root-relative depth fails to integrate information related to the scale of instances. DAPL is incorporated to perceive the scale of instances, which can serve as an auxiliary supervision for the regression of root-relative depth. Moreover, the relative depth of other keypoints suffers cumulative errors from the root point. DAPL facilitates to relieve this issue by indirectly supervising the absolute depth of other keypoints. Specifically, we use the perspective camera model to reconstruct estimated 3D poses using the estimated absolute depth, root-relative depth and 2D groundtruth information in camera coordinates:

$$X_i = \frac{(x_i^* - cx^*)\cdot(z_i - \Delta z_i)}{fx^*}, \tag{9}$$

$$Y_i = \frac{(y_i^* - cy^*)\cdot(z_i - \Delta z_i)}{fy^*}, \tag{10}$$

where x_i^*, y_i^*, are the groundtruth x-axis and y-axis coordinates for the position i in the 2D image plane, cx^*, cy^* are the values of x-axis, y-axis principal point of the camera intrinsic matrix, and fx^*, fy^* are the focal lengths of x-axis, y-axis of the camera.

Then we use the normalized ℓ_1 loss to formulate the pixel-wise projection loss:

$$\mathcal{L}_{\mathrm{p}} = \sum_{i\in S} \frac{1}{B_i} \left\| \frac{(x_i^* - cx^*) \cdot [(z_i - \varDelta z_i) - (z_i^* - \varDelta z_i{}^*)]}{fx^*} \right\|_1 + \sum_{i\in S} \frac{1}{B_i} \left\| \frac{(y_i^* - cy^*) \cdot [(z_i - \varDelta z_i) - (z_i^* - \varDelta z_i{}^*)]}{fy^*} \right\|_1 . \tag{11}$$

In DAPL, the 3D projection model is used to map the predicted absolute depth to the camera coordinate system in combination with the groundtruth 2D positions of the person instances, in the form of indirect supervision of the absolute root depth and the root-relative depth. It is worth noted that we use the groundtruth of 2D position to avoid the suboptimal performance of DAPL caused by naturally existed inaccurate estimation in 2D pose. The mechanism of DAPL will adjust inaccurate predictions of absolute depth and relative depth to the correct optimization direction, thus directly optimizing the absolute and relative 3D pose.

Overall Loss For training the proposed single-stage DRM, we formulate the overall loss function $\mathcal{L}$ as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{c}} + \lambda_o \mathcal{L}_{\mathrm{o}} + \lambda_{rz} \mathcal{L}_{\mathrm{rz}} + \lambda_{\varDelta z} \mathcal{L}_{\varDelta \mathrm{z}} + \lambda_p \mathcal{L}_{\mathrm{p}}, \tag{12}$$

where λ_o, λ_{rz}, $\lambda_{\varDelta z}$ and λ_p are hyper-parameters for balancing different loss items. We set λ_o, λ_{rz}, $\lambda_{\varDelta z}$=0.03, λ_p=0.003, which are experimentally validated.

Inference During testing, an image is fed into DRM, to predict the center map, the offset maps, the root depth map and the root-relative depth maps. First, the candidate 2D poses are obtained by performing the NMS process over the center map combined with the offset maps. After that, the root absolute depth and root-relative depth of each candidate instance are obtained from the root depth map and the root-relative depth maps at the 2D position of root point. Then, the absolute depth for all keypoints are obtained by adding up all the root-relative depth to the root absolute depth. Finally, the NMS process is performed over the candidate 2D poses and absolute depth, and preserve at most 20 candidates for one image. Using these candidate results and camera intrinsic matrix, we can reconstruct the 3D pose through the perspective camera model:

$$[X, Y, Z]^T = ZK^{-1} [x, y, 1]^T , \tag{13}$$

where $[X, Y, Z]$ and $[x, y]$ are 3D and 2D coordinates of a keypoint, respectively, and K is the camera intrinsic matrix.

4. Experiments

4.1. Experiment Setup

Datasets We evaluate the proposed DRM for multi-person 3D pose estimation on two popular challenging benchmarks, *i.e.*, MuPoTS-3D [18] and CMU Panoptic [9].

MuCo-3DHP [18] is a multi-person 3D training set comprised by the MPI-INF-3DHP [17] single-person dataset with groundtruth 3D poses from multi-view marker-less motion capture system. We follow SMAP [39] and use 400k images from it for training our DRM. MuPoTS-3D is a testing set consisting of 8,700 challenging images with occlusions, drastic illumination changes, and lens flares in some of the outdoor footage, making it a convincing testbed to inspect the models' generalization capacity. We use it for evaluation as in SMAP [39].

CMU Panoptic [9] is a large-scale dataset captured in the Panoptic studio, offering 3D pose annotations for multiple people engaged in diverse social activities. we follow Zanfir *et al.* [34] and choose two cameras (*i.e.*, 16 and 30), 165k images from different sequences as our training set, and 9,600 images from four activities (*i.e.*, Haggling, Mafia, Ultimatum, Pizza) as our test set.

Implementation Details Our framework is implemented with PyTorch platform. The proposed model is trained on 8 NVIDIA V100 GPUs with the batch size of 8 per GPU. We use the warmup training strategy and the base learning rate is set as 1×10^{-3}. The learning rate will increase to the basic training rate in the first epoch and then linearly decay to 0 in the end. Adam [4] is used for optimization.

We adopt HRNet [26] as the backbone due to its leading performance in dense prediction tasks, *e.g.*, human pose estimation. The backbone is initialized with the ImageNet [5] pre-trained weights. We train two models for 15 epochs on MuCo-3DHP and CMU Panoptic, separately, mixed with COCO [14] dataset. 50% data in each mini-batch is from COCO. Since COCO lacks 3D pose annotations, weights of 3D losses are set to zero when images from COCO are fed. All images are resized to a fixed size 832×512 as the input to our model.

4.2. Experiment on MuPoTS-3D [18] Benchmark

Evaluation Metrics 3DPCK [20] is a 3D extended version of the Percentage of Correct Keypoints (PCK) metric used in 2D HPE evaluation. An estimated keypoint is considered as correct if the distance between the estimation and the ground-truth is within a certain threshold (*i.e.*, 15cm in our experiments). PCK$_{rel}$ measures the relative pose accuracy with root alignment; PCK$_{abs}$ measures the absolute pose accuracy without root alignment; and PCK$_{root}$ only measures the accuracy of root point.

Comparison with State-of-the-Art Models Tab. 1 shows the result comparisons between our proposed DRM and other state-of-the-art methods. Our single-stage approach, achieves 85.1 PCK$_{rel}$ and 41.0 PCK$_{abs}$, which is superior to all bottom-up methods and most top-down methods except Cheng *et al.* [2] for matched people. Note that we achieve 4.6 PCK$_{rel}$ improvement for the relative 3D pose and 2.3 PCK$_{abs}$ improvement for the absolute 3D pose at the root

Table 1. Comparisons on the MuPoTS-3D [18] dataset. All numbers are average values over 20 activities.

	Methods	Matched people				All people	
		$\text{PCK}_{rel}\uparrow$	$\text{PCK}_{abs}\uparrow$	$\text{PCK}_{root}\uparrow$	$\text{AUC}_{rel}\uparrow$	$\text{PCK}_{rel}\uparrow$	$\text{PCK}_{abs}\uparrow$
Top down	CDMP (ResNet-50) [20]	82.5	31.8	**31.0**	**40.9**	81.8	31.5
	HDnet (FPN) [13]	83.7	35.2	-	-	-	-
	HMOR (FPN) [30]	-	-	-	-	**82.0**	**43.8**
	Pandanet (FPN) [1]	-	-	-	-	72.0	-
	3Dpose (HRNet-w32) [2]	**89.6**	**48.0**	-	-	-	-
Bottom up	Xnect [19]	75.8	-	-	-	70.4	-
	SMAP (Hourglass) [39]	**80.5**	**38.7**	**45.5**	**42.7**	**73.5**	**35.4**
Single-stage	**DRM (Ours, HRNet-w32)**	**85.1**	**41.0**	**45.6**	**45.4**	**80.9**	**39.3**

Table 2. Comparisons of using the root point (denoted as “RC”) and 2D center point (denoted as “CC”) as clue to regress the 2D pose, respectively.

Methods	$\text{AP}\uparrow$	$\text{AP}_M\uparrow$	$\text{AP}_L\uparrow$	$\text{AR}\uparrow$	$\text{AR}_M\uparrow$	$\text{AR}_L\uparrow$
RC	63.9	59.8	71.6	70.2	64.4	78.6
CC	**67.2**	**61.8**	**77.1**	**73.0**	**66.3**	**82.6**

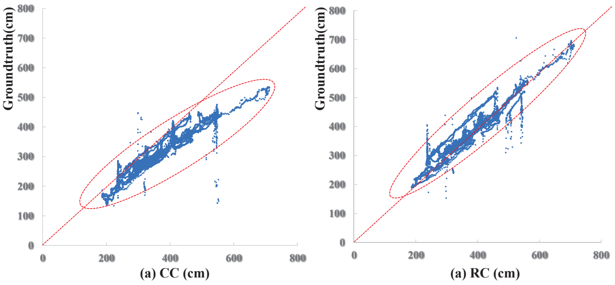

Figure 4. Comparisons of using the root point (denoted as “RC”) and 2D center point (denoted as “CC”) as clue to regress depth information, respectively. The scatter diagrams show the deviation between the predicted absolute depth and corresponding ground truth.

point compared to SMAP [39], which is current the state-of-the-art bottom-up method.

Analysis of the Decoupled Representation Instead of using a single point to encode the all properties of person instance. In this paper, we propose the new decoupled representation that encodes the 2D pose and depth information of each 3D human instance via the center point (center of visible keypoints) and the root point (denoted as pelvis), respectively. We conduct the ablative analysis to explore the superiority of such decoupled representation.

We first employ the the root point (pelvis) and 2D center point to regress x-y keypoint offsets, respectively. The results on COCO dataset [14] are reported in Tab. 2. It can be observed that regressing the 2D pose using the center point performs better than using root point (*i.e.*, 67.2 AP $vs.$ 63.9 AP). Thus, the center point is considered to be able to encode more informative feature, *e.g.*, scale and pose deformation, than the root point.

We further apply the root point (pelvis) and 2D center point to regress the corresponding depth information. The center point’s absolute depth is calculated by averaging the absolute depth of all visible keypoints. As shown in Fig. 4, it can be observed that regressing the depth using root point information achieves less deviation with corresponding ground truth depth, especially for large depth. Thus, the root point is considered to be a position with explicit semantic information than the center point, which is beneficial for depth estimation.

Table 3. Analysis of the Proposed Components. $PDQM$ denotes the 2D Pose-guided Depth Query Module. $DAPL$ indicates the Decoupled Absolute Pose Loss.

$PDQM$	$DAPL$	$\text{PCK}_{abs}\uparrow$	$\text{PCK}_{root}\uparrow$	$\text{PCK}_{rel}\uparrow$
		32.1	32.3	81.3
√		35.5	40.8	81.4
	√	39.8	44.1	83.7
√	√	**41.0**	**45.6**	**85.1**

Therefore, we propose the decoupled representation leveraging different points to encode and predict different properties, *e.g.*, 2D pose and depth, which remarkably improves the estimation of absolute 3D poses.

Analysis of the Proposed Components Based on the decoupled representation, we study the contributions of the two key components in DRM, *i.e.*, 2D Pose-guided Depth Query Module (PDQM) and Decoupled Absolute Pose Loss (DAPL).

As shown in Tab. 3, DAPL obtains the improvement of 11.8 PCK_{root} and 7.7 PCK_{abs}, showing that DAPL significantly promotes the absolute pose estimation by enhancing the absolute depth prediction. Moreover, incorporating PDQM can independently boost the performance by 8.5 PCK_{root} and 3.4 PCK_{abs}, indicating that the scale information in the 2D pose regression branch refines the depth estimation, alleviating poor perception ability on depth. Finally, our full model containing both the DAPL and PDQM, achieves the whole gains of 13.3 PCK_{root}, 8.9 PCK_{abs}, and 3.8 PCK_{rel}, respectively.

Qualitative Result Fig. 5 gives the visualized results of the estimated 3D poses upon in-the-wild images from COCO [14] validation set. It is shown that even in outdoor challenging scenario (containing scale variance, crowds, oc-

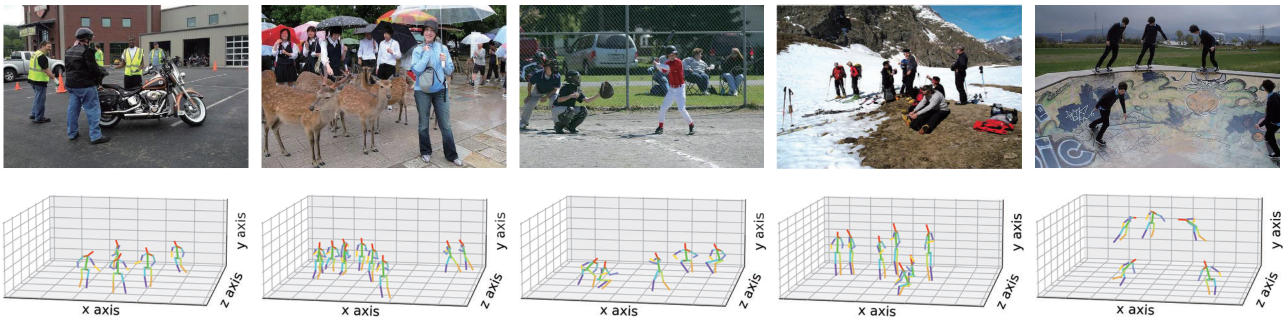

Figure 5. Visualized results of the proposed DRM upon in-the-wild images from COCO [14] validation set. Top row: input images. Bottom row: corresponding multi-person 3D pose estimation results of the proposed DRM.

Table 4. Quantitative comparisons of RtError on CMU Panoptic.

Methods	Haggling	Mafia	Ultim.	Piazza	Mean↓
MPSM [35]	257.8	257.8	301.1	294.0	315.5
CDMP [20]	160.2	151.9	177.5	127.7	154.3
SMAP [39]	84.7	87.7	91.2	78.5	85.5
DRM (Ours)	**63.7**	**58.5**	**52.3**	**69.1**	**60.9**

Table 5. Running time (ms) comparisons.

Methods		3-person↓	20-person↓
CDMP [20] (Top-down)	DetectNet	120.0	120.0
	PoseNet	14.7	71.8
	RootNet	13.0	58.9
	Total	147.7	250.7
SMAP [39] (Bottom-up)	SSNet	57.0	57.0
	Grouping	4.5	8.8
	RefineNet	0.80	0.83
	Total	62.3	66.6
DRM (Ours)	Single-stage	**55.6**	**56.0**

clusion, and huge depth variance), our method still performs surprisingly well.

4.3. Experiment on CMU Panoptic [9] Benchmark

We use RtError [39] and Mean Per Joint Position Error (MPJPE) [34] as the evaluation metrics on CMU Panoptic [9]. RtError measures the absolute estimation of root point and MPJPE measures the accuracy of the 3D root-relative pose. Quantitative comparisons of RtError are provided in Tab. 4. It can be observed that our model significantly outperforms the state-of-the-art bottom-up method SMAP [39] in terms of RtError by a large margin, *i.e.*, 24.6mm improvement (mean of the four activities), showing the promising potential of the proposed DRM for generalization ability. For completeness, we provide the quantitative comparisons of MPJPE between state-of-the-art methods and ours in the supplementary material.

4.4. Running Time Analysis

Tab. 5 reports the detailed comparisons on running time during inference of the representative top-down, bottom-up methods [20,39], and proposed DRM. The experiment is conducted on one NVIDIA V100 GPU. The existing top-down and bottom-up methods both take multi-stage paradigms, leading to computational redundancy. Specifically, the top-down method CDMP [20] adopts a detector to select every single instance, and the total computation cost linearly grows with the number of people. The bottom-up method SMAP [40] needs an additional grouping process to group keypoints to its corresponding instance. In contrast, our single-stage model DRM costs less running time, which hardly increases with the instance number. It is noted that DRM spends 6.7ms less in the 3-person setting and 9.4ms less in the 20-person setting than that of SMAP [39].

5. Conclusion

In this paper, we propose an efficient single-stage Decoupled Regression Model (DRM) to address multi-person absolute 3D pose estimation. DRM utilizes parallel branches to regress 2D pose and human depth simultaneously, which enables a more compact pipeline. Moreover, DRM introduces the 2D Pose-guided Depth Query Module (PDQM) and Decoupled Absolute Pose Loss (DAPL) to jointly advance the accuracy of depth prediction. The PDQM concatenates the features from the 2D pose regression branch to enrich the features for absolute depth regression, which significantly helps to achieve better performance on 3D pose. DAPL maps the predicted depth to the camera coordinate system using the groundtruth 2D position of instances, achieving direct pose supervision in 3D space, which advances the performance of depth prediction. In a further step, we will dedicate to explore more applications of our PDQM to other single-stage methods, *e.g.*, BMP [35] and ROMP [29] for body mesh estimation.

Acknowledgments

The paper is under the funding of National Nature Fund No.62071056, No.62102039 and No.62006244, and Young Elite Scientist Sponsorship Program of China Association for Science and Technology YESS20200140.

References

[1] A. Benzine, F. Chabot, B. Luvison, Q. C. Pham, and C. Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *CVPR*, 2020. 1, 2, 7

[2] Yu Cheng, Bo Wang, Bo Yang, and Robby T. Tan. Monocular 3d multi-person pose estimation by integrating top-down and bottom-up networks. In *CVPR*, pages 7649–7659, 2021. 2, 6, 7

[3] Yu Cheng, Bo Yang, Bo Wang, and Robby T. Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *AAAI*, 34(07):10631–10638, 2020. 1

[4] Kingma D and Ba J. Adam: A method for stochastic optimization. *Computer Science*, 2014. 6

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 6

[6] Hao-Shu Fang, Jinkun Cao, Yu-Wing Tai, and Cewu Lu. Pairwise body-part attention for recognizing human-object interactions. In *ECCV*, 2018. 1

[7] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, 2018. 1

[8] Fang Zhao Xuecheng Nie Yunpeng Chen Shuicheng Yan Jian ZHAO, Jianshu Li and Jiashi Feng. Marginalized cnn: Learning deep invariant representations. In *BMVC*, pages 127.1–127.12, 2017. 1

[9] H. Joo, T. Simon, X. Li, H. Liu, L. Tan, L. Gui, S. Banerjee, T. Godisart, B. Nabbe, and I. Matthews. Panoptic studio: A massively multiview system for social interaction capture. *TPAMI*, pages 1–1, 2016. 2, 6, 8

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

[11] Jianshu Li, Jian Zhao, Congyan Lang, Yidong Li, Yunchao Wei, Guodong Guo, Terence Sim, Shuicheng Yan, and Jiashi Feng. Multi-human parsing with a graph-based generative adversarial model. In *ACMMM*, 2020. 1

[12] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 1

[13] Jiahao Lin and Gim Hee Lee. Hdnet: Human depth estimation for multi-person camera-space localization. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV*, pages 633–648, 2020. 1, 2, 7

[14] T. Y. Lin, M. Maire, S. Belongie, J. Hays, and C. L. Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014. 6, 7, 8

[15] Diogo C. Luvizon, David Picard, and Hedi Tabia. Multi-task deep learning for real-time 3d human pose estimation and action recognition. *TPAMI*, 43(8):2752–2764, 2021. 1

[16] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2

[17] 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*, pages 506–516, 2017. 6

[18] D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, and C. Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3DV*, 2018. 2, 6, 7

[19] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, and Christian Theobalt. Xnect: real-time multi-person 3d motion capture with a single rgb camera. *TOG*, 39(4), 2020. 1, 2, 7

[20] G. Moon, J. Y. Chang, and K. M. Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *ICCV*, 2020. 1, 2, 3, 4, 6, 7, 8

[21] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 2

[22] Xuecheng Nie, Jiashi Feng, Jianfeng Zhang, and Shuicheng Yan. Single-stage multi-person pose machines. In *ICCV*, 2019. 2

[23] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *CVPR*, 2018. 1

[24] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 2

[25] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 1

[26] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[27] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 2

[28] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2

[29] Yu Sun, Qian Bao, Wu Liu, Yili Fu, and Tao Mei. Centerhmr: a bottom-up single-shot method for multi-person 3d mesh recovery from a single image. *ArXiv*, abs/2008.12272, 2020. 8

[30] Can Wang, Jiefeng Li, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *ECCV*, pages 242–259, 2020. 1, 2, 7

[31] Yabo Xiao, Xiaojuan Wang, Dongdong Yu, Guoli Wang, Qian Zhang, and Mingshu He. Adaptivepose: Human parts as adaptive points. In *AAAI*, 2022. 2

[32] Yabo Xiao, Dongdong Yu, Xiaojuan Wang, Lei Jin, Guoli Wang, and Qian Zhang. Learning quality-aware representation for multi-person pose regression. In *AAAI*, 2022. 2

[33] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 1, 2

[34] 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*, pages 2148–2157, 2018. 6, 8

[35] Jianfeng Zhang, Dongdong Yu, Jun Hao Liew, Xuecheng Nie, and Jiashi Feng. Body meshes as points. In *CVPR*, pages 546–556, 2021. 3, 8

[36] Jian Zhao, Jianshu Li, Yu Cheng, Terence Sim, Shuicheng Yan, and Jiashi Feng. Understanding humans in crowded scenes: Deep nested adversarial learning and a new benchmark for multi-human parsing. In *ACMMM*, 2018. 1

[37] Jian Zhao, Jianshu Li, Hengzhu Liu, Shuicheng Yan, and Jiashi Feng. Fine-grained multi-human parsing. *IJCV*, 128, 2020. 1

[38] Jian Zhao, Jianshu Li, Xuecheng Nie, Fang Zhao, Yunpeng Chen, Zhecan Wang, Jiashi Feng, and Shuicheng Yan. Self-supervised neural aggregation networks for human parsing. In *CVPRW*, pages 1595–1603, 2017. 1

[39] 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*, pages 550–566, 2020. 1, 2, 3, 4, 6, 7, 8

[40] Bin Xiao Zhaoxiang Zhang Jingdong Wang Zigang Geng, Ke Sun. Bottom-up human pose estimation via disentangled keypoint regression. In *CVPR*, 2021. 2, 4, 5, 8

[41] Christian Zimmermann, Tim Welschehold, Christian Dornhege, Wolfram Burgard, and Thomas Brox. 3d human pose estimation in rgbd images for robotic task learning. In *ICRA*, pages 1986–1992, 2018. 1