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

Lei Jin¹, Chenyang Xu¹, Xiaojuan Wang¹,¹, Yabo Xiao¹, Yandong Guo², Xuecheng Nie³, Jian Zhao⁴†
¹Beijing University of Posts and Telecommunications
²OPPO Research Institute
³National University of Singapore
⁴Institute 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.

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-
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\textsubscript{rel} and 2.3 PCK\textsubscript{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 \( P = \{p_m\}_{m=1}^{N} \) in \( I \), where \( N \) denotes the number of persons in \( I \). Assume that there are \( J \) keypoints in a single 3D pose skeleton. The \( m-th \) absolute 3D pose can be formulated as: \( p_{m}^{abs} = \{(x_{m,j}^{abs}, y_{m,j}^{abs}, z_{m,j}^{abs})^T\}_{j=1}^{J} \), where \( (x_{m,j}^{abs}, y_{m,j}^{abs}, z_{m,j}^{abs})^T \) is the \( j-th \) absolute pose in the camera-centered coordinate system, as shown in Fig. 2 (c).

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

\[
\begin{align*}
  p_m &= \{(x_{m,j}, y_{m,j})^T\}_{j=1}^{J}, \\
  \Delta Z_m &= \{Z_{m,r} - Z_{m,j}\}_{j=1}^{J},
\end{align*}
\]

where \( (x_{m,j}, y_{m,j})^T \) is the \( j-th \) keypoint position of the \( m-th \) 2D pose in the pixel coordinates, and \( Z_{m,j} \) is the \( j-th \) keypoint absolute depth of the \( m-th \) 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 \( C \) and \( n \) offset maps \( O_m \) 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 \( C^* \). 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 \( I \), \( C^*(x, y) = \exp \left(-\frac{\| (x, y) - (x^c, y^c) \|^2}{\sigma^2}\right) \), where \( (x^c, y^c) \) is the position of an instance center, and \( \sigma \) is the Gaussian variance. Each of the offset maps \( O_m \) 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 \( O^* \) for each image are constructed from all the 2D poses \( \{p_1, p_2, \ldots, p_n\} \) in the image. We compute the center position \( \overline{p}_q = \frac{1}{\pi} \sum_{k=1}^{n} p_k \). The candidate area is around the center position and its radius is

---

![Figure 2. Visualization and explanation of the pose representation.](image)

(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 for the instance.
set to 3 according to previous method [40]. The pixels in the candidate area represent the offset to the keypoints \( \{ \overline{p}_1 - p_1, \overline{p}_2 - p_2, \ldots, \overline{p}_n - p_{in} \} \).

**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 \( Z^* \) 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 \( \Delta Z^* \) 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 \( X \). Then \( X \) 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 \( X \) output from the backbone into \( n \) feature maps, \( \{X_1, X_2, \cdots, X_n\} \), and estimate the offset map \( \{O_i\}; i = 1, 2, \cdots, n \) for each keypoint for the corresponding feature map:

\[
O_i = F_i(X_i), \quad i = 1, 2, \cdots, n, \tag{3}
\]

where \( F_i(\cdot) \) is the \( i \)-th regressor of the \( i \)-th keypoint, and \( O_i \) is the offset map for the \( i \)-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:

\[
Z = F_Z \left( \text{cat} \{X, W(O_1), W(O_2) \cdots W(O_{n-1})\} \right), \tag{4}
\]

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

\[
\mathcal{L}_c = \|C - C^\ast\|^2_2, \tag{5}
\]

where \( \| \cdot \|^2_2 \) is the entry-wise 2-norm, \( C \) and \( C^\ast \) 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 \( \ell_1 \) loss for the dense offset maps:

\[
\mathcal{L}_o = \sum_{i \in \mathcal{S}} \frac{1}{B_i} \text{smooth}_1(o_i - o_i^\ast), \tag{6}
\]

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 \( o_i^\ast \) 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 \( \ell_1 \) loss to formulate the pixel-wise depth loss:

\[
\mathcal{L}_{re} = \sum_{i \in \mathcal{S}} \text{smooth}_1(z_i - z_i^\ast), \tag{7}
\]

\[
\mathcal{L}_{\Delta z} = \sum_{i \in \mathcal{S}} \text{smooth}_1(\Delta z_i - \Delta z_i^\ast), \tag{8}
\]

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

\[
X_i = \frac{(x_i^\ast - cx^\ast) \cdot (z_i - \Delta z_i)}{fx^\ast}, \tag{9}
\]

\[
Y_i = \frac{(y_i^\ast - cy^\ast) \cdot (z_i - \Delta z_i)}{fy^\ast}, \tag{10}
\]

where \( x_i^\ast, y_i^\ast \), are the groundtruth \( x \)-axis and \( y \)-axis coordinates for the position \( i \) in the 2D image plane, \( cx^\ast, cy^\ast \) are the values of \( x \)-axis, \( y \)-axis principal point of the camera intrinsic matrix, and \( fx^\ast, fy^\ast \) are the focal lengths of \( x \)-axis, \( y \)-axis of the camera.
Then we use the normalized $\ell_1$ loss to formulate the pixel-wise projection loss:

$$L_p = \sum_{i \in S} \frac{1}{B_i} \left( \left( x_i^* - cx^* \right) \cdot \left((z_i - \Delta z_i) - (z_i^* - \Delta z_i^*) \right) \right)_{\ell_1} + \sum_{i \in S} \frac{1}{B_i} \left( \left( y_i^* - cy^* \right) \cdot \left((z_i - \Delta z_i) - (z_i^* - \Delta z_i^*) \right) \right)_{\ell_1}. \quad (11)$$

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 $L$ as follows:

$$L = L_s + \lambda_o L_o + \lambda_r L_r + \lambda_{x\Delta} L_{x\Delta} + \lambda_p L_p, \quad (12)$$

where $\lambda_o$, $\lambda_r$, $\lambda_{x\Delta}$, and $\lambda_p$ are hyper-parameters for balancing different loss items. We set $\lambda_o = 0.03$, $\lambda_r = 0.003$, 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:

$$[X, Y, Z]^T = K^{-1} [x, y, 1]^T,$$

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 $1 \times 10^{-3}$. 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.
We conduct the ablative analysis to explore the superiority of the decoupled representation. Analysis of the Decoupled Representation

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Matched people</th>
<th>All people</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PCK&lt;sub&gt;ref&lt;/sub&gt;↑</td>
<td>PCK&lt;sub&gt;rel&lt;/sub&gt;↑</td>
</tr>
<tr>
<td>Top down</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDMP (ResNet-50) [20]</td>
<td>82.5</td>
<td>31.8</td>
</tr>
<tr>
<td>DRM (Ours, HRNet-w32)</td>
<td>78.6</td>
<td></td>
</tr>
<tr>
<td>Bottom up</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xnect [19]</td>
<td>75.8</td>
<td>-</td>
</tr>
<tr>
<td>SMAP (Hourglass) [39]</td>
<td>80.5</td>
<td>38.7</td>
</tr>
<tr>
<td>Single-stage</td>
<td>85.1</td>
<td>41.0</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>AP↑</th>
<th>AP&lt;sub&gt;rel&lt;/sub&gt;↑</th>
<th>AP&lt;sub&gt;root&lt;/sub&gt;↑</th>
<th>AR↑</th>
<th>AR&lt;sub&gt;rel&lt;/sub&gt;↑</th>
<th>AR&lt;sub&gt;root&lt;/sub&gt;↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>RC</td>
<td>63.9</td>
<td>59.8</td>
<td>71.6</td>
<td>70.2</td>
<td>64.4</td>
<td>78.6</td>
</tr>
<tr>
<td>CC</td>
<td>67.2</td>
<td>61.8</td>
<td>77.1</td>
<td>73.0</td>
<td>66.3</td>
<td>82.6</td>
</tr>
</tbody>
</table>

Figure 4. Comparisons of using the root point (denoted as “RC”) and 2D center point (denoted as “CC”) as clue to regress the 2D pose, respectively.

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.

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<sub>root</sub> and 7.7 PCK<sub>abs</sub>, 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<sub>root</sub> and 3.4 PCK<sub>abs</sub>, 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<sub>root</sub>, 8.9 PCK<sub>abs</sub>, and 3.8 PCK<sub>rel</sub>, respectively.

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

<table>
<thead>
<tr>
<th>PDQM</th>
<th>DAPL</th>
<th>PCK&lt;sub&gt;abs&lt;/sub&gt;↑</th>
<th>PCK&lt;sub&gt;root&lt;/sub&gt;↑</th>
<th>PCK&lt;sub&gt;rel&lt;/sub&gt;↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>32.1</td>
<td>32.3</td>
<td>81.3</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>35.5</td>
<td>40.8</td>
<td>81.4</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>39.8</td>
<td>44.1</td>
<td>83.7</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>41.0</td>
<td>45.6</td>
<td>85.1</td>
</tr>
</tbody>
</table>

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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Haggling</th>
<th>Mafia</th>
<th>Ultim.</th>
<th>Piazza</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>MPSM [35]</td>
<td>257.8</td>
<td>257.8</td>
<td>301.1</td>
<td>294.0</td>
<td>315.5</td>
</tr>
<tr>
<td>CDMP [20]</td>
<td>160.2</td>
<td>151.9</td>
<td>177.5</td>
<td>127.7</td>
<td>154.3</td>
</tr>
<tr>
<td>SMAP [39]</td>
<td>84.7</td>
<td>87.7</td>
<td>91.2</td>
<td>78.5</td>
<td>85.5</td>
</tr>
<tr>
<td>DRM (Ours)</td>
<td>63.7</td>
<td>58.5</td>
<td>52.3</td>
<td>69.1</td>
<td>60.9</td>
</tr>
</tbody>
</table>

Table 5. Running time (ms) comparisons.

<table>
<thead>
<tr>
<th>Methods</th>
<th>3-person</th>
<th>20-person</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDMP [20] (Top-down)</td>
<td>DetectNet</td>
<td>120.0</td>
</tr>
<tr>
<td></td>
<td>PoseNet</td>
<td>14.7</td>
</tr>
<tr>
<td></td>
<td>RootNet</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>147.7</td>
</tr>
<tr>
<td>SMAP [39] (Bottom-up)</td>
<td>SSNet</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>Grouping</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>RefineNet</td>
<td>0.80</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>62.3</td>
</tr>
<tr>
<td>DRM (Ours)</td>
<td>Single-stage</td>
<td>55.6</td>
</tr>
</tbody>
</table>

conclusion, 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


