HMOR: Hierarchical Multi-Person Ordinal Relations for Monocular Multi-Person 3D Pose Estimation

Can Wang$^{\star 2}$, Jiefeng Li$^{\star 1}$, Wentao Liu^{2}, Chen Qian^{2}, and Cewu Lu^{1**}

[1] Shanghai Jiao Tong University, Shanghai, China
{ljf_likit, lucewu}@sjtu.edu.cn
[2] SenseTime Research, Beijing, China
{wangcan, liuwentao, qianchen}@sensetime.com

Abstract. Remarkable progress has been made in 3D human pose estimation from a monocular RGB camera. However, only a few studies explored 3D multi-person cases. In this paper, we attempt to address the lack of a global perspective of the top-down approaches by introducing a novel form of supervision - *Hierarchical Multi-person Ordinal Relations (HMOR)*. The HMOR encodes interaction information as the ordinal relations of depths and angles hierarchically, which captures the *body-part* and *joint* level semantic and maintains global consistency at the same time. In our approach, an integrated top-down model is designed to leverage these ordinal relations in the learning process. The integrated model estimates human bounding boxes, human depths, and root-relative 3D poses simultaneously, with a coarse-to-fine architecture to improve the accuracy of depth estimation. The proposed method significantly outperforms state-of-the-art methods on publicly available multi-person 3D pose datasets. In addition to superior performance, our method costs lower computation complexity and fewer model parameters.

Keywords: 3D human pose, ordinal relations, integrated model

1 Introduction

Estimating 3D human poses from a monocular RGB camera is fundamental and challenging. It has found applications in robotics [13, 72], activity recognition [50, 32], human-object interaction detection [15, 51, 28, 29], and content creation for graphics [4, 1]. With deep neural networks [57, 19, 43, 44] and large scale publicly available datasets [56, 21, 3, 31, 23, 36, 38, 33], significant improvement has been achieved in the field of 3D pose estimation. Most of the works [47, 35, 59, 71, 17, 60, 64, 9] focus on estimating the single-person pose. Recently, some methods [52,

[⋆] Denotes equal contribution, order determined by coin flip.

[⋆⋆] Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

38, 67, 66, 53, 48, 39] start to deal with multi-person cases. However, recovering absolute 3D poses in the camera-centered coordinate system is quite a challenge. Since multi-person activities take place in cluttered scenes, inherent depth ambiguity and occlusions make it still difficult to estimate the absolute position of multiple instances.

Recently, top-down approaches [52, 53, 39] achieve noticeable improvements in estimating multi-person 3D poses. These approaches first perform human detection and estimate the 3D pose of each person by a single-person pose estimator. However, the pose estimator is applied to each bounding box separately, which raises the doubt that the top-down models are not able to understand multi-person relationships and handle complex scenes. Without a broad view of the input scenario, it is challenging to get rid of inherent depth ambiguity and occlusion problems. In this paper, the relationship among multiple persons is fully considered to address this limitation of top-down approaches.

We propose a novel form of supervision for 3D pose estimation - *Hierarchical Multi-person Ordinal Relations (HMOR)*. HMOR explicitly encodes the interaction information as ordinal relations, supervising the networks to output 3D poses in the correct order. Different from previous works [46, 61, 54] that only use relative depth information, HMOR considers both depths and angles relations and expresses the ordinal information hierarchically, i.e., *instance* → *part* → *joint*, which makes up for the lack of a global perspective of the top-down approaches.

Further, we propose an integrated top-down model to learn this knowledge by encoding it into the learning process. The integrated model can be end-to-end trained with back-propagation and performs *human detection*, *pose estimation*, and *human-depth estimation* simultaneously. Since metric depth from a single image is fundamentally ambiguous, estimating absolute 3D pose suffers from inaccurate human-depth estimation. To improve the accuracy, we take a coarse-to-fine approach to estimate human depth: i) initializes a global depth map, and ii) finetunes the human depths by estimating the correction residual.

We evaluate our method on two multi-person [38, 23] and one single-person 3D pose datasets [21]. Our method significantly outperforms previous multi-person 3D pose estimation methods [52, 38, 67, 37, 26, 39] by **12.3** PCK$_{abs}$ improvement on the MuPoTS-3D [38] dataset, and **20.5** mm improvement on CMU Panoptic [23] dataset, with lower computation complexity and fewer model parameters. Compared to state-of-the-art single-person methods [17, 59, 60, 68], our method does not need ground-truth bounding-box in the inference phase and still achieves comparable performance. Additionally, our proposed method is compatible with 2D pose annotations, which allows the 2D-3D mixed training strategy.

The contributions of this paper can be summarized as follows:

- We propose HMOR, a novel form of supervision, to explicitly leverage the relationship among multiple persons for pose estimation. HMOR divides human relations into three levels: *instance*, *part* and *joint*. This hierarchical manner ensures both the global consistency and the fine-grained accuracy of the predicted results.

- An integrated end-to-end top-down model is proposed for multi-person 3D pose estimation from a monocular RGB input. We design a coarse-to-fine architecture to improve the accuracy of human-depth estimation. Our model jointly performs human detection, human-depth estimation, and 2D/3D pose estimation.

2 Related Work

Multi-person 2D Pose Estimation. Most of the multi-person 2D pose estimation methods can be divided into two categories: bottom-up and top-down approaches. Bottom-up approaches localize the body joints and group them into different persons. Traditional top-down approaches first detect human bounding boxes in the image and then estimate single-person 2D poses separately.

Representative works [5, 42, 25, 22] of the bottom-up approaches are reviewed. Cao et al. [5] propose part affinity fields (PAFs) to model human bones. Complete skeletons are assembled by detected joints with PAFs. Newell et al. [42] introduce a pixel-wise tag to assign joints to a specific person. Kocabas et al. [25] assign joints to detected persons by a pose residual network.

Top-down approaches [16, 18, 10, 63, 62, 27, 58] achieve impressive accuracy in multi-person 2D pose estimation. Mask R-CNN [18] is an end-to-end model to estimate multiple human poses but still process multiple persons separately. Fang et al. [16] propose a two-stage framework (RMPE) to reduce the effect of the inaccurate human detector. Sun et al. [58] propose the HRNet that maintains high-resolution representations through the whole process.

Single-person 3D Pose Estimation. There are two approaches to the problem of single-person 3D pose estimation from monocular RGB: single-stage and two-stage approaches.

Single-stage approaches [47, 36, 59, 24, 60] directly locate 3D human joints from the input image. For example, Pavlakos et al. [47] propose a coarse-to-fine approach to estimate a 3D heatmap for pose estimation. Kanazawa et al. [24] recover 3D pose and body mesh by minimizing the reprojection loss. Sun et al. [60] operate an integral operation as soft-argmax to obtain 3D pose coordinates in a differentiable manner.

Two-stage approaches [2, 45, 65, 7, 40, 35, 71, 17, 64] first estimate 2D pose or utilize the off-the-shelf accurate 2D pose estimator, and then lift them to the 3D space. Martinez et al. [35] propose a simple baseline to regress 3D pose from 2D coordinates directly. Moreno-Noguer [40] obtains more precise pose estimation by the distance matrix representation. Yang et al. [64] utilize a multi-source discriminator to generate anthropometrically valid poses.

Multi-person 3D Pose Estimation. A few works explore the problem of multi-person 3D pose estimation from a monocular RGB. Rogez et al. [52, 53] propose LCR-Net and LCR-Net++. They locate human bounding boxes and classify those boxes into a set of K anchor-poses. A regression module is proposed to refine the anchor-pose to the final prediction. Instead of using a learning-based manner, they obtain the human depth by minimizing the distance between the projected 3D pose and the estimated 2D pose. Mehta et al. [38] propose a bottom-up method. Their proposed occlusion-robust pose-map (ORPM) enables full body pose inference even under strong partial occlusions. Zanfir et al. [67] propose MubyNet, a bottom-up model. MubyNet integrates a limb scoring model and formulates the person grouping problem as an integer program. Moon et al. [39] propose a top-down two-stage model. They utilize the off-the-shelf human detection model and then perform single-person 3D pose estimation and root-joint localization. Those top-down approaches are not able to utilize multi-person relations since they estimate individual 3D pose separately. The bottom-up approaches are still suffering from limited accuracy. Our method combines the advantages of both approaches and boosts multi-person absolute 3D pose estimation by leveraging the multi-person relations in the integrated end-to-end top-down model.

Ordinal Relations. In the context of computer vision, several works learn ordinal apparent depth [73, 8] or reflectance [41, 69] relationship as weak supervision. They motivated by the fact that ordinal relations are easier for humans to annotate. In the case of single-person 3D pose estimation, [46, 54, 55] use depth relations of body joints to generate 3D pose from 2D pose.

3 Method

We propose a novel representation, *Hierarchical Multi-person Ordinal Relation (HMOR)*, to explicitly leverage ordinal relations among multiple persons and improve the performance of 3D pose estimation. Compared with previous works [46, 61, 54] that use ordinal relation in 3D pose estimation, HMOR extends this idea in three dimensions: i) single-person to multi-persons, ii) *joint* level to hierarchical *instance-part-joint* levels, iii) depth relations to angle relations. Further, we develop an integrated model to aggregate HMOR into the end-to-end training process. In this section, we first describe the unified representation of the absolute multi-person 3D pose recovery under the top-down framework (§3.1). Then we detail the encoding and training schemes of the proposed HMOR (§3.2). Finally, the integrated model with a coarse-to-fine depth estimation design is elaborated (§3.3).

3.1 Representation

Our task is to recover multiple absolute 3D human poses $\mathcal{P} = \{\mathbf{P}_m^{abs}\}_{m=1}^{N}$ in the camera-centered coordinate system, where N denotes the number of persons in the input RGB image. We assume that there are J joints in a single 3D pose skeleton. The m^{th} absolute 3D pose can be formulated as:

$$\mathbf{P}_m^{abs} = \{\mathbf{k}_{m,j} : (x_{m,j}^{abs}, y_{m,j}^{abs}, z_{m,j}^{abs})^\top\}_{j=1}^{J}, \tag{1}$$

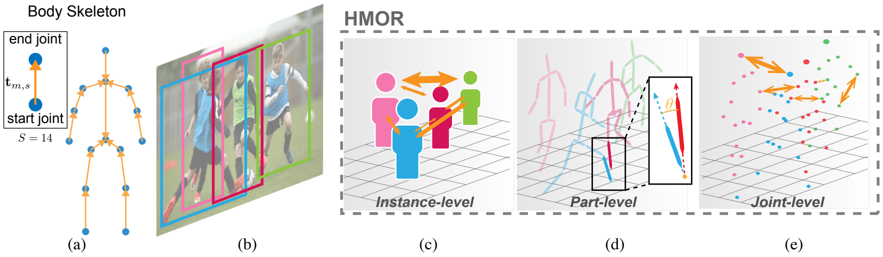

Fig. 1. Illustration of the proposed HMOR. (a) Definition of skeletal parts. (b) Monocular input image. (c-e) Hierarchical Multi-person Ordinal Relations. HMOR supervises the ordinal relations among multiple persons

where $\mathbf{k}_{m,j}$ is the j^{th} joint position of the m^{th} absolute pose.

Human bounding boxes $\{\hat{\mathbf{B}}_m\}_{m=1}^{N}$, root-relative 3D poses $\{\hat{\mathbf{P}}_m^{rel}\}_{m=1}^{N}$, and absolute depth of the root-joint $\{\hat{z}_{m,R}^{abs}\}_{m=1}^{N}$ are needed to estimate the absolute 3D poses. We term root-joint's absolute depth as human depth, corresponding to the pelvis bone position (the R^{th} joint of the body skeleton). We use $\hat{}$ to denote the predicted values. The m^{th} human bounding box $\hat{\mathbf{B}}_m$ and root-relative 3D pose $\hat{\mathbf{P}}_m^{rel}$ are formulated as:

$$\hat{\mathbf{B}}_m = (\hat{u}_m^{top}, \hat{v}_m^{top}, \hat{w}_m, \hat{h}_m)^{\mathsf{T}}, \tag{2}$$

$$\hat{\mathbf{P}}_m^{rel} = \{(\hat{u}_{m,j}, \hat{v}_{m,j}, \hat{z}_{m,j}^{rel})^{\mathsf{T}}\}_{j=1}^{J}, \tag{3}$$

where $\hat{u}_{m,j}$ and $\hat{v}_{m,j}$ represent pixel coordinates of the estimated body joint with respect to the bounding box. $\hat{z}_{m,j}^{rel}$ denotes the estimated depth of joint j relative to the root-joint. $\hat{u}_m^{top}$, $\hat{v}_m^{top}$, $\hat{w}_m$, and $\hat{h}_m$ are the top left corner coordinates, the width, and the height of the predicted bounding box, respectively. With the intrinsic matrix $\mathbf{M}$, the final absolute 3D pose $\hat{\mathbf{P}}_m^{abs}$ can be obtained via back-projection, where each joint is calculated by:

$$\begin{pmatrix} \hat{x}_{m,j}^{abs} \\ \hat{y}_{m,j}^{abs} \\ \hat{z}_{m,j}^{abs} \end{pmatrix} = (\hat{z}_{m,j}^{rel} + \hat{z}_{m,R}^{abs})\mathbf{M}^{-1} \begin{pmatrix} \hat{u}_{m,j} + \hat{u}_m^{top} \\ \hat{v}_{m,j} + \hat{v}_m^{top} \\ 1 \end{pmatrix}. \tag{4}$$

3.2 Hierarchical Multi-person Ordinal Relations

Our initial goal is to leverage multi-person interaction relations to improve the performance of 3D pose estimation. Traditional top-down methods [52, 53, 39] lack a global perspective because they estimate single human poses in each bounding box separately. Therefore, they are vulnerable to truncation, self-occlusions, and inter-person occlusions. Here, we develop a novel form of supervision named *Hierarchical Multi-person Ordinal Relations (HMOR)* to model human relations explicitly. Basically, given an image of human activities, we divide the relationship into three levels: i) instance-level depth relations, ii) part-level angle relations, iii) joint-level depth relations. In each level, HMOR formulates pair-wise ordinal relations and punishes the wrong-order pairs. In the following, we detail our HMOR formulations that reflect interpretable relations of human activities.

Instance-Level Depth Relations. In a given camera view, for two persons $(\mathbf{p}_1, \mathbf{p}_2)$, we denote the instance depth-relation function as $R_{ins}(\mathbf{p}_1, \mathbf{p}_2; \mathbf{n}_\perp)$, taking the value:

- $+1$, if $\mathbf{p}_1$ is closer than $\mathbf{p}_2$ in the $\mathbf{n}_\perp$ direction,
- -1, if $\mathbf{p}_2$ is closer than $\mathbf{p}_1$ in the $\mathbf{n}_\perp$ direction,
- 0, if the depths of two person are equal,

where $\mathbf{n}_\perp$ is the camera normal vector. We define the position of a person as the arithmetic mean of its body joints, i.e. $\mathbf{p}_m = \frac{1}{J}\sum_j^J \hat{\mathbf{k}}_{m,j}$. The ordinal error of a pair of instances is denoted as:

$$err_{ins}(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2) = \log(1 + \max(0, R_{ins}(\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2; \mathbf{n}_\perp) * [(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2) \cdot \mathbf{n}_\perp])). \qquad (5)$$

This differentiable instance ranking expression will punish the wrong-order instance pairs and ignore the correct results. For example, if $\mathbf{p}_1$ is closer than $\mathbf{p}_2$, and the prediction relation is correct, i.e., $(\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2) \cdot \mathbf{n}_\perp < 0$, the multiplication result will be smaller than 0 and ignored by the maximum operation.

Supervising the instance-level depth relations is to help the network build a global understanding of the input scenario. Ablative study in §4.4 reveals that the accuracy of human-depth estimation benefits a lot from instance-level depth relations.

Part-Level Angle Relations. As shown in Fig. 1(a), we divide the body skeleton into $S = 14$ parts according to the kinematically connected joints. Each part $\mathbf{t}$ is a vector defined by start-joint $\mathbf{k}_{start}$ and end-joint $\mathbf{k}_{end}$, i.e., $\mathbf{t} = \mathbf{k}_{end} - \mathbf{k}_{start}$. Since body-parts are a set of 3D vectors with direction and length values, we can not directly compare their depths. Here, we utilize a unique attribute of body-part – direction, and compare their angle relations. To simplify the ordinal relation of angles, we first project the body-part vector $\mathbf{t}_{m,s}$ onto the camera plane:

$$\mathbf{t}_{m,s}^{\mathbf{n}_\perp} = \mathbf{t}_{m,s} - (\mathbf{t}_{m,s} \cdot \mathbf{n}_\perp)\mathbf{n}_\perp, \qquad (6)$$

where m is the person index, and s is the body-part index. In a given camera view, for a pair of body parts $(\mathbf{t}_{m_1,s_1}, \mathbf{t}_{m_2,s_2})$, we denote the angle-relation function as $R_{arg}(\mathbf{t}_{m_1,s_1}, \mathbf{t}_{m_2,s_2}; \mathbf{n}_\perp)$, taking the value:

- $+1$, if $\mathrm{Arg}(\mathbf{t}_{m_1,s_1}^{\mathbf{n}_\perp}) < \mathrm{Arg}(\mathbf{t}_{m_2,s_2}^{\mathbf{n}_\perp})$,
- -1, if $\mathrm{Arg}(\mathbf{t}_{m_1,s_1}^{\mathbf{n}_\perp}) > \mathrm{Arg}(\mathbf{t}_{m_2,s_2}^{\mathbf{n}_\perp})$,
- 0, if $\mathrm{Arg}(\mathbf{t}_{m_1,s_1}^{\mathbf{n}_\perp}) = \mathrm{Arg}(\mathbf{t}_{m_2,s_2}^{\mathbf{n}_\perp})$,

where $\mathrm{Arg}(\mathbf{t}^{\mathbf{n}_\perp})$ computes the principal value of the argument of the projection vector. The ordinal error of a pair of body-parts is:

$$err_{part}(\hat{\mathbf{t}}_{m_1,s_1}, \hat{\mathbf{t}}_{m_2,s_2}) = [R_{arg}(\hat{\mathbf{t}}_{m_1,s_1}, \hat{\mathbf{t}}_{m_2,s_2}; \mathbf{n}_\perp) * [(\hat{\mathbf{t}}_{m_1,s_1} \times \hat{\mathbf{t}}_{m_2,s_2}) \cdot \mathbf{n}_\perp]]_+. \quad (7)$$

With the cross-product operation ×, we supervise the direction of the angle between a pair of body-parts. If the angle between $\hat{\mathbf{t}}_{m_1,s_1}$ and $\hat{\mathbf{t}}_{m_2,s_2}$ is in the correct direction, the projection of the cross-product $(\hat{\mathbf{t}}_{m_1,s_1} \times \hat{\mathbf{t}}_{m_2,s_2}) \cdot \mathbf{n}_\perp$ will have an opposite sign of $R_{arg}(\cdot)$. Therefore, the negative multiplication results will be ignored by the $[\cdot]_+$ operation.

Another intuitive way is to express body-parts as particles and supervise their depth relations, using the average position of its two endpoints. To compare vector and particle representations, we conduct ablative experiments and find out vector is superior to particle representation. We suspect this is because the depth relations have been fully utilized in the other two levels, supervising depths of body-part is redundant. More experimental details are reported in §4.4.

Joint-Level Depth Relations. The definition of body joint depth-relation function $R_{jt}(\mathbf{k}_{m_1,s_1}, \mathbf{k}_{m_2,s_2}; \mathbf{n}_\perp)$ is similar to R_{ins}:

- $+1$, if $\mathbf{k}_{m_1,s_1}$ is closer than $\mathbf{k}_{m_2,s_2}$ in the $\mathbf{n}_\perp$ direction,
- -1, if $\mathbf{k}_{m_2,s_2}$ is closer than $\mathbf{k}_{m_1,s_1}$ in the $\mathbf{n}_\perp$ direction,
- 0, if the depths of two joints are equal.

The ordinal error of a pair of joints is denoted as:

$$err_{jt}(\hat{\mathbf{k}}_{m_1,s_1}, \hat{\mathbf{k}}_{m_2,s_2}) = \log(1 + [R_{jt}(\hat{\mathbf{k}}_{m_1,s_1}, \hat{\mathbf{k}}_{m_2,s_2}; \mathbf{n}_\perp]_+ * [(\hat{\mathbf{k}}_{m_1,s_1} - \hat{\mathbf{k}}_{m_2,s_2}) \cdot \mathbf{n}_\perp])). \quad (8)$$

Denoting the set of estimated *persons*, *body-parts*, and *joints* pairs as $\mathcal{I}_{ins}$, $\mathcal{I}_{part}$, and $\mathcal{I}_{jt}$, respectively, the HMOR loss is computed as follows:

$$\mathcal{L}_{HMOR} = \frac{1}{|\mathcal{I}_{ins}|} \sum_{\hat{\mathbf{p}}_1, \hat{\mathbf{p}}_2} err_{ins} + \frac{1}{|\mathcal{I}_{part}|} \sum_{\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_2} err_{part} + \frac{1}{|\mathcal{I}_{jt}|} \sum_{\hat{\mathbf{k}}_1, \hat{\mathbf{k}}_2} err_{jt} \quad (9)$$

Augmented Training Scheme. As mentioned before, HMOR computes the ordinal relations with respect to a vector $\mathbf{n}_\perp$. Initially, this vector is set as the camera normal vector. However, we notice that annotations from 3D human pose datasets (Human3.6M, MuPoTS-3D, and CMU Panoptic) are mostly captured in an laboratory environment, limited to the fixed viewing angle. To alleviate camera restrictions, we sample virtual views to improve the generalization ability.

In the training phase, we generate a virtual view vector $\mathbf{n}_v$ by rotating the camera normal vector $\mathbf{n}_\perp$ randomly. We adapt the uniform sphere sampling strategy from Marsaglia et al. [34]:

$$\mathbf{n}_v = (\sqrt{1-u^2}\cos\theta, \sqrt{1-u^2}\sin\theta, u)^\mathsf{T}, \quad (10)$$

where $\theta \sim U[0, 2\pi)$ and $u \sim U[0, 1]$. In this way, HMOR can calculate the ordinal relations with respect to an arbitrary viewing angle. The effectiveness of the sampled view is validated in §4.4.

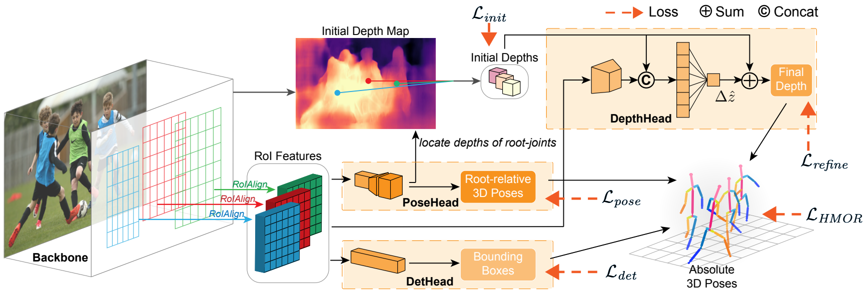

Fig. 2. Architecture of the integrated model. The ResNet-50 based backbone network extract RoI features and initial depth map. PoseHead and DetHead perform root-relative 3D pose estimation and human detection, respectively. DepthHead retrieves initial depths from the depth map and predicts refined human depths by correction residual $\Delta\hat{z}$. This architecture allows the 2D-3D mixed training strategy

Additionally, a mixed datasets training strategy is utilized for a fair comparison with previous methods in experiments. HMOR is compatible with 2D pose datasets and single-person 3D pose datasets. Given an image only with 2D pose annotations, we can define the part-level angle relations, since the 2D pose skeletons are the projections of body-parts with respect to $\mathbf{n}_{\perp}$. As for single-person cases, HMOR only supervises the *joint* and *body-part* relations of an individual person and ignore instance-level relations.

3.3 Integrated End-to-end Model

In our approach, an integrated end-to-end top-down model is designed to aggregate HMOR into the end-to-end training process. Although the disjoint model [39] can use different strong networks for different tasks (e.g., human detection, pose estimation, depth estimation), an integrated model has three advantages over the disjoint learning model: 1) Fewer model parameters. 2) Only an integrated model can leverage the multi-person relations since the disjoint learning methods train their model with single person annotations separately. 3) The multi-task training strategy of the integrated model can benefit each task. In our experiments, the integrated model is found to have much better performance than the disjoint learning methods.

The overall architecture of our model is summarized in Fig. 2. Our model consists of two stages. In the first stage, the backbone network extracts RoIs and the initial depth map. PoseHead and DetHead estimate root-relative 3D poses and human bounding boxes from RoIs, respectively. In the second stage, we retrieve the initial depths of root-joints from the depth map. The DepthHead takes the RoI features and initial depth as input and outputs the correction residual $\Delta\hat{z}$. The residual is added to the initial depth to obtain refined human depths. Aggregating the outputs from DetHead, PoseHead, and DepthHead, the absolute 3D poses $\hat{\mathbf{P}}^{abs}$ are estimated via back-projection as Eq. 4.

Human Detection. The architecture for human detection and the loss function L_{det} follow the design in Mask R-CNN [18]. Region Proposal Network (RPN) proposes candidate human bounding boxes, and the DetHead predicts class labels and bounding-box offsets. RoiAlign is used to extract feature maps from each RoI.

Root-Relative 3D Pose Estimation. PoseHead is proposed to estimate the root-relative 3D pose $\hat{\mathbf{P}}^{rel}$ from an input RoI feature. We use 3D heatmaps as the representations of 3D poses. The soft-argmax operation [60] is adopted to extract $\hat{\mathbf{P}}^{rel}$ from the 3D heatmap. ℓ_1 regression loss is applied to root-relative coordinates $\hat{\mathbf{P}}^{rel}_m$:

$$\mathcal{L}_{pose} = \frac{1}{N}\frac{1}{J}\sum_{m}^{N} \|\hat{\mathbf{P}}^{rel}_m - \mathbf{P}^{rel}_m\|_1. \tag{11}$$

RoIAlign extracts 14 × 14 RoI features, which are fed into the PoseHead subsequently. We adopt a simple network as PoseHead, including three residual blocks for feature extraction, a transposed convolution [14] for upsampling, a batch normalization layers[20], a ReLU activation function, and a 1×1 convolution. The size of an output heatmap is $28 \times 28 \times 28$.

Human Depth Estimation. Direct human-depth regression from an input RoI is challenging. Part of the challenges comes from the variety of camera parameters and human body shapes. Furthermore, the inputs of DepthHead are fixed-size RoI features, which erase the information of projected body shapes and sizes. Inspired by the idea of iterative error feedback (IEF) from previous works [6, 12, 24], we design a coarse-to-fine estimation approach to enhance the accuracy of human-depth regression. The model will first predict an initial depth of root-joint $\hat{z}^{init}$. Then the DepthHead takes the RoI features and the initial depth $\hat{z}^{init}$ as an input and outputs the residual Δz. Ideally, the refined depth is updated by adding this residual to the initial estimate $\hat{z}^{refine} = \hat{z}^{init} + \Delta z$.

Depth Initialize. To estimate the initial depths of root-joints, we directly regress an initial depth map. During training, we first normalize the absolute depth value by focal lengths and then calculate the loss $\mathcal{L}_{init}$ between the ground truth and the initial depth map in the area around the root-joint's 2D pixel location:

$$z^{norm}_R = z^{abs}_R / \sqrt{f_x \cdot f_y}. \tag{12}$$

$$\mathcal{L}_{init} = \frac{1}{N}\sum_{m}^{N} \|z^{norm}_{m,R} - \hat{z}^{init}_{m,R}\|_1, \tag{13}$$

Depth Refinement. In the refinement step, we retrieve the initial-depth values of root-joints from the depth map according to their 2D pixel locations. Because the input features are resized by RoIAlign, we first need to transfer the original depth to the equivalent depth of the resized person. According to the pinhole camera model:

$$z_R^{eq,norm} = z_R^{norm} \cdot \sqrt{\frac{A_{Box}}{A_{RoI}}}, \tag{14}$$

$$\hat{z}_R^{eq,init} = \hat{z}_R^{init} \cdot \sqrt{\frac{A_{Box}}{A_{RoI}}}, \tag{15}$$

where A_{Box} denotes the area of the bounding box, and A_{RoI} denotes the area of RoI. DepthHead extracts 1D features from RoIs. Then the equivalent initial-depth values $\hat{z}_{m,R}^{eq,init}$ are concatenated with the extracted features and fed into an *fc* layer to predict the residual $\Delta\hat{z}$. The loss function of the refinement step $\mathcal{L}_{refine}$ is defined as:

$$\mathcal{L}_{refine} = \frac{1}{N}\sum_{m}^{N} \|z_{m,R}^{eq,norm} - \hat{z}_{m,R}^{eq,init} - \Delta\hat{z}_m\|_1. \tag{16}$$

In the testing phase, we can recover the absolute depth of root-joint $\hat{\mathbf{z}}_{m,R}^{abs}$ as:

$$\hat{\mathbf{z}}_{m,R}^{abs} = (\Delta\hat{z}_m + \hat{\mathbf{z}}_{m,R}^{eq,init}) \cdot \sqrt{\frac{f_x \cdot f_y \cdot A_{RoI}}{A_{Box}}}. \tag{17}$$

The DepthHead uses three residual blocks (following ResNet [19]) and an average pooling layer to extract 1D features. The FC layer contains 512 neurons.

The end-to-end training loss is formulated as:

$$\mathcal{L} = \mathcal{L}_{det} + \mathcal{L}_{pose} + \mathcal{L}_{init} + \mathcal{L}_{refine} + \mathcal{L}_{HMOR}. \tag{18}$$

4 Experiment

In this section, we first introduce the datasets employed for quantitative evaluation and elaborate implementation details. Then we report our results and compare the proposed method with state-of-the-art methods. Finally, ablation experiments are conducted to evaluate our contributions and show how each choice contributes to our state-of-the-art performance.

4.1 Datasets

MuCo-3DHP and MuPoTS-3D: MuCo-3DHP is a multi-person composited 3D human pose training dataset. MuPoTS-3D is the real-world scenes test set. Following [38, 39], 400K composited frames are utilized for training.

CMU Panoptic. CMU Panoptic [23] is a multi-person 3D pose dataset captured in an indoor dome with multiple cameras. Here we follow the evaluation protocol of [66, 67].

3DPW. 3D Poses in the Wild (3DPW) [33] is a recent challenging dataset, captured mostly in outdoor conditions. It contains 60 video sequences (24 train, 24 test, and 12 validation).

Human3.6M. Human3.6M [21] is an indoor benchmark for single-person 3D pose estimation. A total of 11 professional actors (6 male, 5 female) perform 15 activities in a laboratory environment.

Table 1. Quantitative comparisons with state-of-the-art methods on the MuPoTS-3D dataset. "-" shows the results that are not available

Method	$\text{AUC}_{rel}\uparrow$	$\text{3DPCK}_{rel}\uparrow$	$\text{3DPCK}_{abs}\uparrow$
LCRNet [52]	-	53.8	-
Single Shot [38]	-	66.0	-
LCRNet++ [53]	-	70.6	-
Xnect [37]	-	70.4	-
Moon et al. [39]	39.8	81.8	31.5
Ours	**43.5**	**82.0**	**43.8**

Table 2. Quantitative comparisons of MPJPE on the CMU Panoptic dataset

Method	Haggling	Mafia	Ultimatum	Pizza	Mean↓
Popa [49]	217.9	187.3	193.6	221.3	203.4
Zanfir [66]	140.0	165.9	150.7	156.0	153.4
Zanfir [67]	72.4	78.8	66.8	94.3	72.1
Ours	**50.9**	**50.5**	**50.7**	**68.2**	**51.6**

4.2 Implementation Details

Our method is implemented in PyTorch. We adopt a ResNet-50 [19] based FPN [30] as our model backbone. The backbone is initialized with the ImageNet [11] pre-trained model. The settings of each network head are reported in §3.3. We resize the image to 1333 × 800 and feed into the network. SGD is used for optimization, with a mini-batch size of 32. All tasks are trained simultaneously. We adopt the linear learning rate warm-up policy. The learning rate is set to 0.2/3 at first and gradually increases to 0.2 after 2.5k iterations. We reduce the learning rate by a factor of 10 at the 10th and 20th epochs. In each experiment, our model is trained for 30 epochs with 16 NVIDIA 1080 Ti GPUs. We perform data augmentations including horizontal flip and multi-scale training. Additional COCO [31] 2D pose estimation data are used in the training phase. For evaluation, we report the flip-test results. All reported numbers have been obtained with a single model without ensembling.

4.3 Compare with Prior Art

MuPoTS-3D. We compare our method against state-of-the-art methods under three protocols. PCK_{abs} is used to evaluate absolute camera-centered coordinates of 3D poses. Additionally, PCK_{rel} and AUC_{rel} are used to evaluate root-relative 3D poses after root alignment. Quantitative results are reported in Table 1. Without bells and whistles, our method surpasses state-of-the-art meth-

Table 3. Quantitative comparisons on the Human3.6M dataset

	Single-Person							*Multi-Person*		
Method	Moreno	Zhou	Martinez	Sun	Fang	Sun	Zhou	Rogez	Moon	**Ours**
	[40]	[70]	[35]	[59]	[17]	[60]	[68]	[53]	[39]	
PA MPJPE↓	76.5	55.3	47.7	48.3	45.7	40.6	**27.9**	42.7	35.2	**30.5**

	Single-Person					*Multi-Person*				
Method	Martinez	Fang	Sun	Sun	Zhou	Rogez	Metha	Rogez	Moon	**Ours**
	[35]	[17]	[59]	[60]	[68]	[52]	[38]	[53]	[39]	
MPJPE↓	62.9	60.4	59.1	49.6	**39.9**	87.7	69.9	63.5	54.4	**48.6**

ods by 12.3 PCK$_{abs}$ (**39.0**% relative improvement). Our method demonstrates a clear advantage for handling multi-person 3D poses.

As for root-relative results, our method achieves 82.0 PCK$_{rel}$ and 43.5 AUC$_{rel}$. Note that the PCK result relies on the threshold value. AUC can reflect a more reliable result since it computes the area under the PCK curve from various thresholds. Our method outperforms the previous methods by 3.7 AUC$_{rel}$.

CMU Panoptic. Following previous works [66, 67], we evaluate our method under MPJPE after root alignment. Table 2 provides experimental results. In this dataset, the activities take place in a small room. Thus, the scenarios are severely affected by the occlusion problem. Our method effectively reduces the interference of occlusion and outperforms state-of-the-art methods by 20.5 mm MPJPE (**28.4**% relative improvement).

Human3.6M. We conduct experiments on Human3.6M dataset to evaluate the performance of root-relative 3D pose estimation. Two experimental protocols are widely used. *Protocol 1* uses *PA MPJPE* and *Protocol 2* uses *MPJPE* as evaluation metrics. As most of the previous methods use the ground-truth bounding box, our method does not require any ground-truth information at inference time. Quantitative results are reported in Table 3. Our method achieves comparable performance with single-person methods and outperforms previous multi-person 3D pose estimation methods.

4.4 Ablation Study

In this study, we evaluate the effectiveness of the proposed HMOR and integrated model. We evaluate on 3DPW dataset that contains in-the-wild complex scenes to demonstrate the strength of our model. We further propose ABS-MPJPE to evaluate the absolute 3D pose estimation results without root alignment.

Effect of Hierarchical Multi-person Ordinal Relations. In this experiment, we study the effectiveness of using HMOR supervision. We first implement a vanilla baseline without HMOR supervision. Moreover, we implement another baseline by directly supervising the predicted absolute 3D poses with an ℓ_1 loss

Table 4. Ablative study on the effects of HMOR

	Settings	3DPW		
		MPJPE↓	PA- ↓	ABS- ↓
(a)	baseline	95.7	63.6	169.3
	+ $\mathcal{L}_{abs}$	94.6	61.1	158.2
	+ *jt*	89.9	59.7	132.8
	+ *part*	90.2	60.3	143.2
	+ *instance*	93.3	61.2	128.3
	+ *jt* + *part*	89.1	58.3	125.9
	+ *jt* + *instance*	89.2	58.5	122.3
	+ *part* + *instance*	89.5	59.5	123.6
	+ *jt* + *part* + *instance*	88.3	57.8	119.6
(b)	+ *jt* + *particle-part* + *instance*	89.0	58.2	119.5
(c)	+ *jt* + *part* + *instance* + *sample views* (Final)	**87.7**	**57.4**	**118.5**
(d)	w/o refine depth	88.4	58.1	133.6

$\mathcal{L}_{abs}$. Intuitively, since the human poses are evaluated in the camera coordinate system, the local optimum for $\mathcal{L}_{abs}$ is consistent with the evaluation metrics.

The experimental results are shown in Table 4(a). The model trained with $\mathcal{L}_{abs}$ supervision has better performance than the vanilla baseline, but is still inferior to HMOR supervision. HMOR supervision brings 7.4 mm MPJPE improvement. By removing three types of relations separately, we can observe that *instance* relation affects the absolute pose accuracy (ABS-MPJPE) most, while *part* and *joint* relations mainly affect the root-relative pose accuracy.

Variants of HMOR. In this experiment, we examine a variant of HMOR. When handling the part relations, we represent a body part as a particle rather than a vector. The position of a body part is defined as the average of its two endpoints. Similar to *joint* and *instance*, we supervise the depth relations of the particle body-parts. The experimental results are shown in Table 4(b). The particle representation produces inferior performance than the vector representation.

Effect of Sampled Views. Table 4(c) reports the result of training with sampled views. Compare with the results in Table 4(a) that only use the original camera normal vector $\mathbf{n}_{\perp}$, sampled views provide 0.6 mm MPJPE improvement.

Effect of Coarse-to-Fine Depth Surpervision. In this experiment, we study the effectiveness of the coarse-to-fine design for human depth estimation. We remove the refinement step and output the initial value directly. The experimental results are shown in Table 4(d). We observe that the coarse-to-fine design is necessary to produce accurate human depth.

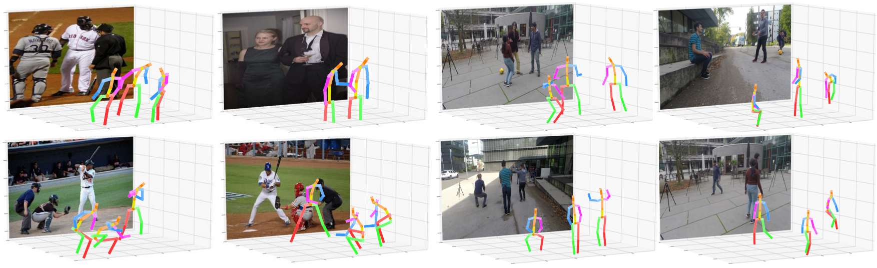

Fig. 3. Qualitative results of our proposed method on COCO validation set (left) and MuPoTS-3D test set (right)

Table 5. Ablative study on computation complexity and model parameters

Method	#Params↓	GFLOPs↓	AUC$_{rel}$↑	PCK$_{rel}$↑	PCK$_{abs}$↑
Moon [39]	167.7M	547.8	39.8	81.8	31.5
Ours	**45.0M**	**320.2**	**43.5**	**82.0**	**43.8**

Computation Complexity. The experimental results of computation complexity and model parameters are listed in Table 5. We compare our method with Moon et al. [39], which is the only open-source multi-person 3D pose estimation method. Our approach obtains superior results to the state-of-the-art 3D pose estimation method (both absolute pose and root-relative pose), with significantly lower computation complexity and fewer model parameters.

5 Conclusion

In this paper, we proposed a novel form of supervision - HMOR, to learn multi-person 3D poses from a monocular RGB image. HMOR supervises the multi-person ordinal relations in a hierarchical manner, which captures fine-grained semantics and maintains global consistency at the same time. To end-to-end learn the ordinal relations, we further proposed an integrated model with a coarse-to-fine depth-estimation architecture. We demonstrate the effectiveness of our proposed method on standard benchmarks. The proposed method surpasses state-of-the-art multi-person 3D pose estimation methods, with lower computation complexity and fewer model parameters. We believe the idea of leveraging multi-person relations can be further explored to improve 3D pose estimation, e.g., exploit the relations via network design.

Acknowledgements. This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332Shanghai Qi Zhi Institute.

References

1. Aitpayev, K., Gaber, J.: Creation of 3D human avatar using kinect. ATFECM (2012)
2. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3D human pose reconstruction. In: CVPR (2015)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2D human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
4. Boulic, R., Bécheiraz, P., Emering, L., Thalmann, D.: Integration of motion control techniques for virtual human and avatar real-time animation. In: VRST (1997)
5. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2D pose estimation using part affinity fields. In: CVPR (2017)
6. Carreira, J., Agrawal, P., Fragkiadaki, K., Malik, J.: Human pose estimation with iterative error feedback. In: CVPR (2016)
7. Chen, C.H., Ramanan, D.: 3D human pose estimation = 2D pose estimation + matching. In: CVPR (2017)
8. Chen, W., Fu, Z., Yang, D., Deng, J.: Single-image depth perception in the wild. In: NeurIPS (2016)
9. Chen, X., Lin, K.Y., Liu, W., Qian, C., Lin, L.: Weakly-supervised discovery of geometry-aware representation for 3D human pose estimation. In: CVPR (2019)
10. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018)
11. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
12. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR (2010)
13. Du, G., Zhang, P.: Markerless human–robot interface for dual robot manipulators using kinect sensor. ROBOT CIM-INT MANUF (2014)
14. Dumoulin, V., Visin, F.: A guide to convolution arithmetic for deep learning. arXiv:1603.07285 (2016)
15. Fang, H.S., Cao, J., Tai, Y.W., Lu, C.: Pairwise body-part attention for recognizing human-object interactions. In: ECCV (2018)
16. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)
17. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3D pose estimation. In: AAAI (2018)
18. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017)
19. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
20. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
21. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3.6m: Large scale datasets and predictive methods for 3D human sensing in natural environments. TPAMI (2014)
22. Jin, S., Liu, W., Ouyang, W., Qian, C.: Multi-person articulated tracking with spatial and temporal embeddings. In: CVPR (2019)
23. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015)
24. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)

25. Kocabas, M., Karagoz, S., Akbas, E.: MultiPoseNet: Fast multi-person pose estimation using pose residual network. In: ECCV (2018)
26. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: ICCV (2019)
27. Li, J., Wang, C., Zhu, H., Mao, Y., Fang, H.S., Lu, C.: Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In: CVPR (2019)
28. Li, Y.L., Liu, X., Lu, H., Wang, S., Liu, J., Li, J., Lu, C.: Detailed 2d-3d joint representation for human-object interaction. In: CVPR (2020)
29. Li, Y.L., Xu, L., Liu, X., Huang, X., Xu, Y., Wang, S., Fang, H.S., Ma, Z., Chen, M., Lu, C.: Pastanet: Toward human activity knowledge engine. In: CVPR (2020)
30. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR (2017)
31. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014)
32. Luvizon, D.C., Tabia, H., Picard, D.: Learning features combination for human action recognition from skeleton sequences. Pattern Recognition (2017)
33. von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., Pons-Moll, G.: Recovering accurate 3d human pose in the wild using imus and a moving camera. In: ECCV (2018)
34. Marsaglia, G., et al.: Choosing a point from the surface of a sphere. The Annals of Mathematical Statistics (1972)
35. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3D human pose estimation. In: ICCV (2017)
36. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3D human pose estimation in the wild using improved cnn supervision. In: 3DV (2017)
37. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H.P., Rhodin, H., Pons-Moll, G., Theobalt, C.: Xnect: Real-time multi-person 3d human pose estimation with a single rgb camera. arXiv preprint arXiv:1907.00837 (2019)
38. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3D pose estimation from monocular rgb. In: 3DV (2018)
39. Moon, G., Chang, J.Y., Lee, K.M.: Camera distance-aware top-down approach for 3D multi-person pose estimation from a single rgb image. In: ICCV (2019)
40. Moreno-Noguer, F.: 3D human pose estimation from a single image via distance matrix regression. In: CVPR (2017)
41. Narihira, T., Maire, M., Yu, S.X.: Learning lightness from human judgement on relative reflectance. In: CVPR (2015)
42. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NeurIPS (2017)
43. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
44. Pang, B., Zha, K., Cao, H., Shi, C., Lu, C.: Deep rnn framework for visual sequential applications. In: CVPR (2019)
45. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2D pose information. In: ECCV (2016)
46. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3D human pose estimation. In: CVPR (2018)
47. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3D human pose. In: CVPR (2017)

48. Pirinen, A., Gärtner, E., Sminchisescu, C.: Domes to drones: Self-supervised active triangulation for 3d human pose reconstruction. In: NeurIPS (2019)
49. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2D and 3D human sensing. In: CVPR (2017)
50. Presti, L.L., La Cascia, M.: 3d skeleton-based human action classification: A survey. Pattern Recognition (2016)
51. Qi, S., Wang, W., Jia, B., Shen, J., Zhu, S.C.: Learning human-object interactions by graph parsing neural networks. In: ECCV (2018)
52. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR (2017)
53. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2D and 3D pose detection in natural images. TPAMI (2019)
54. Ronchi, M.R., Mac Aodha, O., Eng, R., Perona, P.: It's all relative: Monocular 3D human pose estimation from weakly supervised data. In: BMVC (2018)
55. Sharma, S., Varigonda, P.T., Bindal, P., Sharma, A., Jain, A.: Monocular 3D human pose estimation by generation and ordinal ranking. In: ICCV (2019)
56. Sigal, L., Balan, A.O., Black, M.J.: Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV (2010)
57. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556 (2014)
58. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
59. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017)
60. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
61. Wang, M., Chen, X., Liu, W., Qian, C., Lin, L., Ma, L.: Drpose3d: Depth ranking in 3D human pose estimation. IJCAI (2018)
62. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018)
63. Xiu, Y., Li, J., Wang, H., Fang, Y., Lu, C.: Pose flow: Efficient online pose tracking. arXiv preprint arXiv:1802.00977 (2018)
64. Yang, W., Ouyang, W., Wang, X., Ren, J., Li, H., Wang, X.: 3D human pose estimation in the wild by adversarial learning. In: CVPR (2018)
65. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3D pose estimation from a single image. In: CVPR (2016)
66. Zanfir, A., Marinoiu, E., Sminchisescu, C.: Monocular 3D pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In: CVPR (2018)
67. Zanfir, A., Marinoiu, E., Zanfir, M., Popa, A.I., Sminchisescu, C.: Deep network for the integrated 3D sensing of multiple people in natural images. In: NeurIPS (2018)
68. Zhou, K., Han, X., Jiang, N., Jia, K., Lu, J.: Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In: ICCV (2019)
69. Zhou, T., Krahenbuhl, P., Efros, A.A.: Learning data-driven reflectance priors for intrinsic image decomposition. In: ICCV (2015)
70. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. TPAMI (2018)

71. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3D human pose estimation in the wild: a weakly-supervised approach. In: ICCV (2017)
72. Zimmermann, C., Welschehold, T., Dornhege, C., Burgard, W., Brox, T.: 3D human pose estimation in rgbd images for robotic task learning. In: ICRA (2018)
73. Zoran, D., Isola, P., Krishnan, D., Freeman, W.T.: Learning ordinal relationships for mid-level vision. In: ICCV (2015)