HDNet: Human Depth Estimation for Multi-Person Camera-Space Localization

Jiahao Lin[0000−0001−6210−9608] and Gim Hee Lee[0000−0002−1583−0475]

Department of Computer Science, National University of Singapore
{jiahao,gimhee.lee}@comp.nus.edu.sg

Abstract. Current works on multi-person 3D pose estimation mainly focus on the estimation of the 3D joint locations relative to the root joint and ignore the absolute locations of each pose. In this paper, we propose the Human Depth Estimation Network (HDNet), an end-to-end framework for absolute root joint localization in the camera coordinate space. Our HDNet first estimates the 2D human pose with heatmaps of the joints. These estimated heatmaps serve as attention masks for pooling features from image regions corresponding to the target person. A skeleton-based Graph Neural Network (GNN) is utilized to propagate features among joints. We formulate the target depth regression as a bin index estimation problem, which can be transformed with a soft-argmax operation from the classification output of our HDNet. We evaluate our HDNet on the root joint localization and root-relative 3D pose estimation tasks with two benchmark datasets, *i.e.*, Human3.6M and MuPoTS-3D. The experimental results show that we outperform the previous state-of-the-art consistently under multiple evaluation metrics. Our source code is available at: https://github.com/jiahaoLjh/HumanDepth.

Keywords: Human Depth Estimation · Multi-person Pose Estimation · Camera Coordinate Space

1 Introduction

Human pose estimation is one of the active research topics in the community of computer vision and artificial intelligence due to its importance in many applications such as camera surveillance, virtual/augmented reality, and human-computer interaction, *etc.* Extensive research has been done for human pose estimation in both 2D image space and 3D Cartesian space, respectively. Great successes have been achieved in the single-person 2D/3D pose estimation tasks thanks to the rapid development of deep learning techniques and the emergence of large-scale human pose datasets [1, 12, 16]. On the other hand, multi-person 2D/3D pose estimation tasks are more challenging due to the unknown number of persons in the scene. To mitigate this problem, the multi-person pose estimation task is typically tackled in a two-stage scheme that decouples the estimation of the number of persons and the pose of each person, *i.e.*, the top-down [3, 5, 8, 10, 28] or bottom-up [2, 22, 24] scheme. In recent years, large-scale

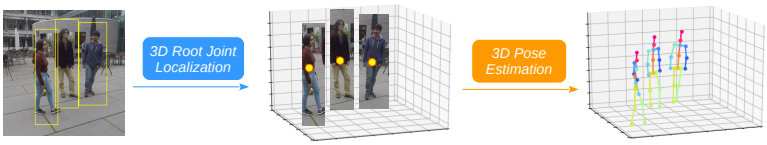

Fig. 1. Top-down multi-person 3D pose estimation pipeline. Camera-space root joint coordinate is estimated for each detected person bounding box, followed by a root-relative 3D pose estimation, to obtain the absolute 3D poses and locations.

multi-person 3D pose datasets such as MuPoTS-3D [20] are created to faciliate the research of multi-person 3D pose estimation. However, most of the existing works [4, 20, 26, 27] focus on the estimation of 3D pose relative to the root joint of each person in the scene. Global absolute locations of the respective 3D poses with respect to the camera coordinate space are ignored.

Estimating the absolute 3D location of each pose in an image is essential for understanding human-to-human interactions. Recently, Moon *et al.* [21] propose a multi-stage pipeline for the task of multi-person 3D pose estimation in the camera coordinate space as shown in Figure 1. The pipeline adopts the top-down scheme which predicts a bounding box for each person in the first stage. This is followed by estimation of the absolute root joint location for the person in each bounding box. Finally, the global pose of each person is recovered by applying single-person 3D pose estimation to get the relative location of other joints with respect to the root joint. The root joint localization framework proposed in [21] estimates the depth of root joint for each person based on the size of the bounding box. Despite showing promising results, the approach relies on the size of bounding box for root joint localization, and hence is not sufficiently effective due to two reasons: (1) The compactness of bounding boxes varies from person to person and also between different object detectors. (2) Sizes of bounding boxes carries no direct information about the size of the particular person due to the variation of poses.

In this paper, we propose an end-to-end Human Depth Estimation Network (HDNet) to address the problems of root joint depth estimation and localization. We adopt the same top-down pipeline for the task of multi-person absolute 3D pose estimation. Our key observation is that we can estimate the depth of a person in a monocular image with considerably high accuracy by leveraging on the prior knowledge of the typical size of the human pose and body joints. Inspired by this observation, we propose to jointly learn the 2D human pose and the depth estimation tasks in our HDNet. More specifically, we utilize the heatmaps of the joints from the human pose estimation task as attention masks to achieve pose-aware feature pooling in each joint type. Subsequently, we put the pose-aware features of the joints into a skeleton-based Graph Neural Network (GNN), where information are effectively propagated among body joints to enhance depth estimation. Following a recent work on scene depth estimation [7], we formulate our depth estimation of the root joint as a classification task, where

the target depths are discretized into a preset number of bins. We also adopt a soft-argmax operation on the bins predicted by our HDNet for faster convergence during training and better performance without losing precision compared to direct numerical depth regression. Our approach outperforms previous state-of-the-art [21] on the task of root joint localization on two benchmark datasets, *i.e.*, Human3.6M [12] and MuPoTS-3D [20] under multiple evaluation metrics. Experimental results also show that accurate root localization benefits the task of root-aligned 3D human pose estimation.

Our contributions in this work are:

- An end-to-end Human Depth Estimation Network (HDNet) is proposed to address the problem of root joint localization for multi-person 3D pose estimation in the camera-space.
- Several key components are introduced in our framework: (1) pose heatmaps are used as attention masks for pose-aware feature pooling; (2) a skeleton-based GNN is designed for effective information propagation among the body joints; and (3) depth regression of the root joint is formulated as a classification task, where the classification output is transformed to the estimated depth with a soft-argmax operation to facilitate accurate depth estimation.
- Quantitative and qualitative results show that our approach consistently outperforms the state-of-the-art on multiple benchmark datasets under various evaluation metrics.

2 Related Works

Human pose estimation has been an interesting yet challenging problem in computer vision. Early methods use a variety of hand-crafted features such as silhouette, shape, SIFT features, HOG for the task. Recently, with the power of deep neural networks and well-annotated large-scale human pose datasets, increasing learning-based approaches are proposed to tackle this challenging problem.

Single-person 2D pose estimation. Early works, such as Stacked Hourglass [23], Convolutional Pose Machines [30], *etc.*, have been proposed to use deep convolutional neural networks as feature extractors for 2D pose estimation. Heatmaps of joints are the commonly used representation to indicate the presence of joints at spatial locations with Gaussian peaks. More recent works including RMPE [5], CFN [10], CPN [3], HRNet [28], *etc.*, introduce various framework designs to improve the joint localization precision.

Single-person 3D pose estimation. Approaches for 3D pose estimation can be generally categorized into two groups. Direct end-to-end estimation of 3D pose from RGB images regresses both 2D joint locations and the z-axis root-relative depth for each joint. [25, 29] extend the notion of heatmap to the 3D space, where estimation is performed in a volumetric space. Another group of approaches decouples the task into a two-stage pipeline. The 2D joint locations

are first estimated, followed by a 2D-to-3D lifting. [6, 18] utilize Multi-Layer Perceptron (MLP) to learn the mapping.

Multi-person 2D pose estimation. Top-down [3, 5, 8, 10, 28] and bottom-up [2, 22, 24] approaches have been proposed to estimate poses for multiple persons. Top-down approaches utilize a human object detector to localize the bounding box, followed by a single-person pose estimation pipeline with image patch cropped from the bounding box. Bottom-up approaches detect human joints in a person-agnostic way, followed by a grouping process to identify joints belonging to the same person. Top-down approaches usually estimate joint locations more precisely because bounding boxes of different sizes are scaled to the same size in the single-person estimation stage. However, top-down approaches tend to be more computationally expensive due to the redundancy in bounding box detections.

Multi-person 3D pose estimation. Several works [4, 20, 26, 27, 31] have been conducted on multi-person 3D pose estimation. Rogez *et al.* [26] propose a LCR-Net which consists of localization, classification, and regression parts and estimates each detected human with a classified and refined anchor pose. Mehta *et al.* [20] propose a bottom-up approach which estimates a specially designed occlusion-robust pose map and readout the 3D poses given 2D poses obtained with Part Affinity Fields [2]. Dabral *et al.* [4] propose to incorporate hourglass network into Mask R-CNN detection heads for better 2D pose localization, followed by a standard residual network to lift 2D poses to 3D. Zanfir *et al.* [31] design a holistic multi-person sensing pipeline, *i.e.* MubyNet, to jointly address the problems of multi-person 2D/3D skeleton/shape-based pose estimation. However, these works only estimate and evaluate the 3D pose after root joint alignment and ignore the global location of each pose. Recently, Moon *et al.* [21] propose a multi-stage pipeline for multi-person camera-space 3D pose estimation. The pipeline follows the top-down scheme and consists of a RootNet which localizes the root joint for each detected bounding box. We also adopt the top-down scheme pipeline and estimate the camera-space root joint location and 3D pose for each detected bounding box. To our best knowledge, [21] and our work are the only two works that focus on the estimation and evaluation of multi-person root joint locations. Compared to [21] which relies on the size of detected bounding box, we utilize the underlying features and design a human-specific pose-based root joint depth estimation framework to significantly boost the root localization performance.

3 Our Approach

3.1 Overview

Given a 2D image with an unknown number of persons, the task of camera-space multi-person 3D pose estimation is to: (1) identify all person instances,

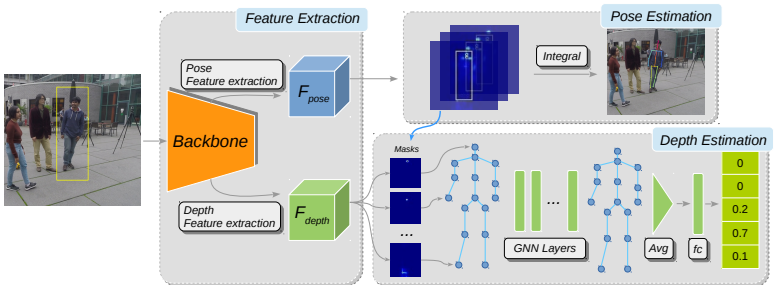

Fig. 2. Our HDNet architecture. The framework takes an image together with the bounding box of a target person as input. A Feature Pyramid Network backbone is used for general feature extraction followed by separated multi-scale feature extraction for the tasks of pose and depth estimation. Estimated heatmaps are used as attention masks to pool depth features. A Graph Neural Network is utilized to propagate and aggregate features for the target person depth estimation.

(2) estimate the 3D pose with respect to the root joint, *i.e.*, pelvis, for each person, and (3) localize each person by estimating the 3D coordinate of root joint in the camera coordinate space.

Following the top-down approaches in the literature of multi-person pose estimation, we assume that the 2D human bounding boxes for each person in the input image are available from a generic object detector. Given the person instances and detected bounding boxes, we propose an end-to-end depth estimation framework to localize the root joint of each person in the camera coordinate space as illustrated in Figure 2. The root joint localization is decoupled into two sub-tasks: (1) localization of the root joint image coordinate (u, v), and (2) estimation of the root joint depth Z in the camera frame, which is then used to back-project (u, v) to 3D space. We use an off-the-shelf single-person 3D pose estimator to estimate the 3D joint locations of each person with respect to the root joint. The final absolute 3D pose of each person in the camera coordinate system is obtained by the transformation of each joint location with the absolute location of the root joint.

The details of our proposed root joint localization framework are introduced in Section 3.2. The choices of specific object detector and single-person 3D pose estimator used in our experiments are given in the implementation details in Section 4.2.

3.2 Root Localization Framework

Our framework for monocular image single/multi-person depth estimation is shown in Figure 2. The framework consists of a Feature Pyramid Network (FPN)-based backbone, a heatmap-based human pose estimation branch, and a Graph Neural Network (GNN)-based depth estimation branch.

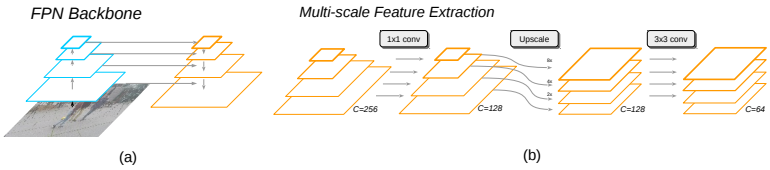

Fig. 3. (a) ResNet-based Feature Pyramid Network Backbone for general feature extraction. (b) Multi-scale feature extraction subnet architecture used for both Pose feature and Depth feature extraction.

Backbone Network. We choose FPN [15] as our backbone network due to its capability of explicitly handling features of multiple scales in the form of feature pyramids. Hence, it is suitable for perceiving the scale of human body parts and consequently enhances depth estimation of the human pose in an image. The FPN network consists of a ResNet-50 [9] with feature blocks of four different scales C_2, C_3, C_4, C_5 (cyan layers in Figure 3(a)), where a reversed hierarchy of feature pyramid P_5, P_4, P_3, P_2 is built upon (orange layers in Figure 3(a)). Each of the four scales encodes hierarchical levels of feature representations, which are then passed through two consecutive convolutional layers as shown in Figure 3(b). An upsampling operation with corresponding upsample scale factor is applied between the two convolutional layers to ensure matching spatial resolution from the output of the four scales. Batch Normalization [11] and ReLU operations are used after each convolution layer. Weights are not shared across scales. Blocks of all scales are then concatenated to form the final feature block $\mathbf{F}$. Since we find that the downstream tasks of pose estimation and depth estimation are not collaboratively correlated, we split the multi-scale feature processing from the output feature pyramid P_5, P_4, P_3, P_2 of the backbone into two parallel branches without shared weights as shown in Figure 2. We denote the features as $\mathbf{F}_{\text{pose}}$ and $\mathbf{F}_{\text{depth}}$, respectively.

2D Pose Estimation Branch. We propose to use estimated 2D pose as a guide to aggregate information from useful feature regions to effectively distil information from the image and discard irrelevant areas such as the background. We first regress N_J heatmaps $\hat{\mathbf{H}}$ that correspond to the N_J joints with a 1×1 convolution from feature block $\mathbf{F}_{\text{pose}}$. Each of the N_J heatmaps are normalized across all spatial locations with a softmax operation. A direct read out of the coordinate from the local maximum limits the precision of the joint location estimation due to the low resolution of the output heatmap (4x downsample from input image in ResNet backbone). To circumvent this problem, we follow the idea of "soft-argmax" in [29] and compute the "integral" version of estimated coordinate $(\hat{u}, \hat{v})$ for each joint j using the weighted sum of coordinates:

$$(\hat{u}_j, \hat{v}_j) = \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \hat{\mathbf{H}}^{(j)}_{u,v} \cdot (u, v), \tag{1}$$

where W and H are the width and height of output heatmap. The softmax operation guarantees that the weights $\hat{\mathbf{H}}_{u,v}$ form a valid distribution which sum up to 1 over all spatial locations. To supervise the heatmap regression, we generate a ground truth heatmap $\mathbf{H}^{(j)\mathrm{GT}}$ for each joint j. A Gaussian peak is created around the ground truth joint location (u_j, v_j) with a preset standard deviation that controls the compactness of the Gaussian peak. We use standard Mean Squared Error (MSE) as the heatmap regression loss and $L1$ loss for the pose after soft-argmax as follows:

$$\mathcal{L}_{\mathrm{hm}} = \frac{1}{N_J HW} \sum_{j}^{N_J} \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \left\| \mathbf{H}_{u,v}^{(j)\mathrm{GT}} - \hat{\mathbf{H}}_{u,v}^{(j)} \right\|^2, \tag{2}$$

$$\mathcal{L}_{\mathrm{pose}} = \frac{1}{N_J} \sum_{j}^{N_J} \left(\left| u_j^{\mathrm{GT}} - \hat{u}_j \right| + \left| v_j^{\mathrm{GT}} - \hat{v}_j \right| \right). \tag{3}$$

To deal with multiple persons in the image, we focus on a target person by zeroing out the regions of the heatmap outside the bounding box of that person from the object detector.

Depth Estimation Branch. After we obtain the heatmaps, we use them as attention masks to guide the network into focusing on specific regions of the image related to the target person. More specifically, we only care about features from pixel locations that are close to the joints of the target person. The intuition behind our design choice is that joint locations contain more scale-related information than the larger yet less discriminative areas such as the whole upper body trunk. Attention-guided feature pooling is also adopted in other tasks such as action recognition [17] and hand pose estimation [13]. We compute the weighted sum feature vector $\mathbf{d}$ for each joint j from the feature block $\mathbf{F}_{\mathrm{depth}}$ as:

$$\mathbf{d}^{(j)} = \sum_{(u,v)=(0,0)}^{(W-1,H-1)} \hat{\mathbf{H}}_{u,v}^{(j)} \cdot \mathbf{F}_{\mathrm{depth}_{u,v}}. \tag{4}$$

To effectively aggregate features corresponding to different joint types, we formulate a standard Graph Neural Network (GNN) where each node represents one joint type, *e.g.*, elbow, knee, *etc.* The aggregated features $\mathbf{d}^{(j)}$ for each joint type j is fed into the corresponding node $X_{in}^{(j)}$ in the graph as input. Each layer of the GNN is defined as:

$$X_{out}^{(i)} = \sigma\Big(\tilde{a}_{ii} f_{self}(X_{in}^{(i)}; \Theta_{self}) + \sum_{j \neq i} \tilde{a}_{ij} f_{inter}(X_{in}^{(j)}; \Theta_{inter}) \Big). \tag{5}$$

The feature of each input node X_{in} undergoes the linear mappings $f_{self}(.)$ and $f_{inter}(.)$ that are parametrized by Θ_{self} and Θ_{inter}, respectively. The output of the node, i.e. $X_{out}^{(i)}$ is computed from a weighted aggregation of $f_{self}(.)$ and

$f_{inter}(.)$ of all other nodes. The weighting factor $\tilde{a}_{ij}$ is an element of the normalized adjacency matrix $\tilde{A} \in \mathbb{R}^{N_j \times N_j}$ that controls the extent of influence of the nodes on each other. The original adjacency matrix is $A \in \{0,1\}^{N_j \times N_j}$; an element a_{ij} equals 1 if there is a skeletal link between joint i and j, *e.g.* left knee to left ankle, or otherwise 0. $\tilde{A} \in \mathbb{R}^{N_j \times N_j}$ is obtained by applying $L1$-normalization on each row of A. The non-linearity function $\sigma(.)$ is implemented with a Batch Normalization followed by a ReLU. We stack L GNN layers in total. After the last GNN layer, we merge the feature output from each node with an average pooling operation.

Target output formulation Inspired by the work [7] for scene depth estimation, we formulate the depth estimation as a classification problem instead of directly regressing the numerical value of depth. We follow the practice in [7] to discretize the log-depth space into a preset number of bins, $N_{\mathbf{B}}$. We compute:

$$b(d) = \frac{\log d - \log \alpha}{\log \beta - \log \alpha} \cdot (N_{\mathbf{B}} - 1), \tag{6}$$

where $\lfloor b \rceil$ gives the bin index of the depth, and the depth d of a pose is assumed to be within the range $[\alpha, \beta]$. Here $\lfloor . \rceil$ is the round-off to the nearest integer operator. To eliminate quantization errors, we assign non-zero values to two consecutive bins i and $i+1$, where $i \leq b < i+1$. This operation is similar to the weights in bi-linear interpolation. For example, the ground truth values of the bins are given by $\mathbf{B} = [0, 0, 0.6, 0.4, 0]$ for $N_{\mathbf{B}} = 5$ and $b = 2.4$. Consequently, $\mathbf{B}$ is a 1D heatmap that can achieve any level of precision with a sufficiently accurate categorical estimation on the bins.

Since a different focal length of the camera affects the scale of a target person in the image, it is unrealistic to estimate the absolute depth from images taken by any arbitrary camera. To alleviate this problem, we normalize out the camera intrinsic parameters by replacing the target d with $\hat{d} = d/f$, where f is the focal length of camera. We approximate with $\hat{d} = d/\sqrt{f_x \cdot f_y}$ in our experiments since the focal lengths in x and y directions are usually very close. Finally, we add a fully connected layer after the pooled feature from the last GNN layer to regress the $N_{\mathbf{B}}$ values of the bins $\hat{\mathbf{B}}$. Softmax operation is used to normalize the output into a valid distribution. We transform $\hat{\mathbf{B}}$ back to the estimated depth d of the root joint by:

$$d = \exp\left[\frac{\hat{b}}{N_{\mathbf{B}} - 1} \cdot (\log \beta - \log \alpha) + \log \alpha\right] \cdot \sqrt{f_x \cdot f_y}, \text{ where } \hat{b} = \sum_{i=0}^{N_{\mathbf{B}}-1} \hat{\mathbf{B}}_i \cdot i. \tag{7}$$

Similar to the soft-argmax operation used to transform heatmaps to joint locations, $\hat{b}$ is the weighted sum of the bin indices with the estimated heatmap $\hat{\mathbf{B}}$. To supervise the learning of the depth estimation branch, we adopt cross-entropy loss on the estimated bins $\hat{\mathbf{B}}$ and $L1$ loss on $\hat{b}$ as follows:

$$\mathcal{L}_{\text{bins}} = -\sum_{i=0}^{N_{\mathbf{B}}-1} \mathbf{B}_i^{\text{GT}} \cdot \log \hat{\mathbf{B}}_i, \text{ and } \mathcal{L}_{\text{idx}} = \left| b^{\text{GT}} - \hat{b} \right|. \tag{8}$$

We train the whole framework with losses from the pose estimation and depth estimation branches:

$$\mathcal{L} = \lambda_{\text{hm}}\mathcal{L}_{\text{hm}} + \lambda_{\text{pose}}\mathcal{L}_{\text{pose}} + \lambda_{\text{bins}}\mathcal{L}_{\text{bins}} + \lambda_{\text{idx}}\mathcal{L}_{\text{idx}}. \tag{9}$$

4 Experiments

4.1 Datasets and Evaluation Metrics

Human3.6M dataset. Human3.6M dataset [12] is currently the largest publicly available dataset for human 3D pose estimation. The dataset consists of 3.6 million video frames captured by MoCap system in a constrained indoor studio environment. 11 actors performing 15 activities are captured from 4 camera viewpoints. 3D ground truth poses in world coordinate system and camera extrinsic (rotation and translation with respect to world coordinate) and intrinsic parameters (focal length and principal point) are available. We follow previous works that five subjects (S1, S5, S6, S7, S8) are used in training and two subjects (S9 and S11) are used for evaluation. We use every 5th and 64th frames in each video for training and evaluation respectively. No extra 2D pose dataset is used to augment the training. We follow the metric Mean Root Position Error (MRPE) proposed in [21] to evaluate the root localization accuracy. Specifically, we consider the Euclidean distance between the estimated and the ground truth 3D coordinate of the root joint.

MuCo-3DHP and MuPoTS-3D datasets. MuCo-3DHP and MuPoTS-3D are two datasets proposed by Mehta *et al.* [20] to evaluate multi-person 3D pose estimation performance. The training set MuCo-3DHP is a composite dataset which merges randomly sampled 3D poses from single-person 3D human pose dataset MPI-INF-3DHP [19] to form realistic multi-person scenes. The test set MuPoTS-3D is a markerless motion captured multi-person dataset including both indoor and outdoor scenes. We use the same set of MuCo-3DHP synthesized images from [21] for a fair comparison. No extra 2D pose dataset is used to augment the training. For evaluation of multi-person root joint localization, we follow [21] to report the average precision and recall of 3D root joint location under different thresholds. A root joint with a smaller distance to the matched ground truth root joint location than a threshold is considered a true positive estimation. We follow [21] to report 3DPCK$_{abs}$ for evaluation of the root-aware 3D pose estimation, where 3DPCK (3D percentage of correct keypoints) for the estimated poses is evaluated without root alignment. 3DPCK treats an estimated joint as correct if it is within 15 cm distance from the matched ground truth joint. Although our framework does not focus on root-relative 3D pose estimation, we also report the root-aligned 3DPCK$_{rel}$ to show that accurate root localization also benefits the precision of 3D pose estimation.

Table 1. MRPE results comparison with state-of-the-arts on the Human3.6M dataset. MRPE_x, MRPE_y, and MRPE_z are the average errors in x, y, and z axes, respectively.

Method	MRPE	MRPE_x	MRPE_y	MRPE_z
Baseline	267.8	27.5	28.3	261.9
Baseline w/o limb joints	226.2	24.5	24.9	220.2
Baseline with RANSAC	213.1	24.3	24.3	207.1
RootNet [21]	120.0	23.3	23.0	108.1
Ours	**77.6**	**15.6**	**13.6**	**69.9**

4.2 Implementation Details

Following previous work [21], we use Mask R-CNN [8] as our person detector due to its high performance and generalizability to in-the-wild images. For single-person 3D pose estimation, we use the volumetric-based 3D pose estimator by [29]. Instead of cropping out areas of interest using bounding boxes, we keep the original scale of image and crop out a fixed size patch centered around the bounding box, or the principal point if no bounding box is provided in the single-person scenario. The cropped out image is then rescaled to 256×256 and used as input to our network. The output resolution of the heatmap is 64×64. We use 2 layers of GNN operations in the depth estimation branch. We set the standard deviation of the Gaussian peak in the ground truth heatmap to be 0.75, and the bin range of d/f to $[\alpha = 1.0, \beta = 8.0]$ for a reasonably sufficient range of the depth. We do not see much performance change when different number of bins $N_{\mathbf{B}}$ are used. All results of the experiments shown in the paper are obtained with $N_{\mathbf{B}} = 71$. We set λ in Eq. 9 to balance the four loss terms to same order of magnitudes. For training, we use Adam optimizer [14] with learning rate 1e-4 and batch size 16. We train the model for 200k steps and decay the learning rate with a factor of 0.8 at every 20k steps. The evaluation of each image takes around 7ms with our root joint localization HDNet.

4.3 Results on Human3.6M

The root joint localization results on Human3.6M dataset are shown in Table 1. The baselines reported in the top 3 rows follow a two-stage approach, where 2D pose [29] and 3D pose [18] are estimated separately, and an optimization process is adopted to obtain the global root joint location that minimizes the reprojection error. "w/o limb joints" refers to optimization using only head and body trunk joints. "with RANSAC" refers to randomly sampling the set of joints used for optimization with RANSAC. The baseline results are taken from the figures reported in [21]. We also compare with the state-of-the-art approach [21]. It can be seen from Table 1 that optimization-based methods can achieve reasonable results, but with limited accuracy due to the errors from both the 2D and 3D estimation stages. Our root joint localization framework achieves 69.9mm

Table 2. Root joint localization accuracy comparison in average precision and recall with state-of-the-arts on MuPoTS-3D dataset.

Method	AP_{25}^{root}	AP_{20}^{root}	AP_{15}^{root}	AP_{10}^{root}	AR_{25}^{root}	AR_{20}^{root}	AR_{15}^{root}	AR_{10}^{root}
RootNet [21]	31.0	21.5	10.2	2.3	55.2	45.3	31.4	15.2
Ours	**39.4**	**28.0**	**14.6**	**4.1**	**59.8**	**50.0**	**35.9**	**19.1**

Table 3. Sequence-wise $3DPCK_{abs}$ comparison with state-of-the-arts on MuPoTS-3D dataset. Accuracy is measured on matched ground-truths.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	-
RootNet [21]	**59.5**	**45.3**	51.4	**46.2**	**53.0**	**27.4**	23.7	26.4	39.1	23.6	-
Ours	21.4	22.7	**58.3**	27.5	37.3	12.2	**49.2**	**40.8**	**53.1**	**43.9**	-
Method	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
RootNet [21]	18.3	14.9	38.2	**29.5**	36.8	23.6	14.4	20.0	18.8	25.4	31.8
Ours	**43.2**	**43.6**	**39.7**	28.3	**49.5**	**23.8**	**18.0**	**26.9**	**25.0**	**38.8**	**35.2**

depth estimation error in $MRPE_z$ with a 35% improvement over [21]. Since our approach uses the original scale image without scaling to person bounding box size which limits the 2D (u, v) localization precision, we also adopt a state-of-the-art 2D pose estimator CPN [3] within the person bounding box area to further refine the (u, v) localization. Our MRPE for root joint achieves an overall performance of 77.6mm which significantly outperforms the state-of-the-art.

4.4 Results on MuPoTS-3D

Root Joint Localization. To evaluate our root joint localization performance on the multi-person MuPoTS-3D dataset, we estimate the root joint 3D coordinate for each bounding box detected from the object detector. All root joint candidates are matched with the ground truth root joints, and only candidates with distance to the matched ground truth lesser than a threshold are considered as an accurate estimate. We then analyze the average precision and recall over the whole dataset under various settings of thresholds ranging from 25cm to 10cm. The results are shown in Table 2. Our method achieves much higher AP and AR consistently across all levels of thresholds compared to the state-of-the-art approach [21].

Camera-space absolute 3D pose estimation. We also evaluate the camera-space absolute 3D pose estimation performance with $3DPCK_{abs}$. $3DPCK_{abs}$ compares the estimated 3D pose with the matched ground truth pose in the

Table 4. Joint-wise 3DPCK$_{abs}$ comparison with state-of-the-arts on MuPoTS-3D dataset. Accuracy is measured on matched ground-truths.

Method	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Avg
RootNet [21]	37.6	35.6	34.0	34.1	30.7	30.6	31.3	25.3	31.8
Ours	**38.3**	**37.8**	**36.2**	**37.4**	**34.0**	**34.9**	**36.4**	**29.2**	**35.2**

Table 5. Sequence-wise 3DPCK$_{rel}$ comparison with state-of-the-arts on MuPoTS-3D dataset. Accuracy is measured on matched ground-truths.

Method	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	-
Rogez *et al.* [26]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	-
Mehta *et al.* [20]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	-
Rogez *et al.* [27]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	-
RootNet [21]	**94.4**	78.6	79.0	82.1	86.6	72.8	**81.9**	75.8	**90.2**	90.4	-
Ours	**94.4**	**79.6**	**79.2**	**82.4**	**86.7**	**73.0**	81.6	**76.3**	90.1	**90.5**	-
Method	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	Avg
Rogez *et al.* [26]	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Mehta *et al.* [20]	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	**90.9**	70.8
Rogez *et al.* [27]	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
RootNet [21]	**79.4**	**79.9**	75.3	81.0	**81.1**	90.7	**89.6**	83.1	81.7	77.3	82.5
Ours	77.9	79.2	**78.3**	**85.5**	**81.1**	**91.0**	88.5	**85.1**	**83.4**	90.5	**83.7**

camera coordinate space without root alignment, thus requires highly accurate root joint localization. We use the same 3D pose estimator [29] as the state-of-the-art root joint localization method [21] for a fair comparison. Results in Table 3 show that our method consistently outperforms the state-of-the-art in most of the test sequences and achieves a 35.2% average 3DPCK (3.4% improvement). The performance breakdown of all joint types is shown in Table 4.

Root-relative 3D pose estimation. The state-of-the-art root-relative 3D pose estimator [29] adopts a volumetric output representation and estimates the root-relative depth for each joint. Absolute root joint depth has to be available to recover the 3D pose through back-projection. We follow [21] and use our estimated root depth to back-project the 3D pose and evaluate the root-relative 3D pose estimation accuracy with 3DPCK$_{rel}$ after root joint alignment. Results are shown in Table 5, where our method outperforms the previous best performance by 1.2% average 3DPCK. This demonstrates that more accurate root localization also benefits the precise 3D pose estimation in volumetric-based approaches [21, 25, 29].

Table 6. Ablation studies on components of the framework. Depth error MRPE_z (mm) on Human3.6M dataset and $\text{AP}_{25}^{\text{root}}$ (%) on MuPoTS-3D dataset are measured.

Method	$\text{MRPE}_z(\downarrow)$	$\text{AP}_{25}^{\text{root}}(\uparrow)$
RootNet [21]	108.1	31.0
Ours direct regression	94.5	27.3
Ours shared feature branch	72.0	31.9
Ours w/o GNN	72.9	32.7
Ours w/o HM pooling	71.8	26.0
Ours (full)	**69.9**	**39.4**

4.5 Ablation Studies

We conduct ablation studies to show how each component in our framework affects the root joint localization accuracy. We evaluate the depth estimation accuracy MRPE_z on Human3.6M dataset and the root joint localization $\text{AP}_{25}^{\text{root}}$ on MuPoTS-3D dataset with different variants of our framework in Table 6. The state-of-the-art approach [21] is also included for comparison.

- "Ours direct regression": Performance drop (by 24.6mm and 12.1%) with directly regressing target depth instead of performing classification over binning shows the effectiveness of formulating the depth estimation as a classification task.
- "Ours shared feature branch": One single multi-scale feature branch is kept after FPN, which means $\mathbf{F}_{\text{pose}}$ and $\mathbf{F}_{\text{depth}}$ use the same feature representation. This setting causes performance to drop (by 2.1mm and 7.5%), and thus demonstrates that the features used for pose estimation and depth estimation are not highly correlated.
- "Ours w/o GNN": We replace the GNN layers in our depth estimation branch with same number of fully-connected layers and observe a performance drop (by 3mm and 6.7%), showing the effectiveness of the graph neural network in propagating and refining the features extracted for different types of joints.
- "Ours w/o HM pooling": We remove feature pooling with estimated heatmaps as mask in the depth estimation branch and instead apply a global average pooling to obtain a single feature vector. The GNN layers are replaced with fully-connected layers since we do not explicitly differentiate between different joint types. We observe a performance drop (by 1.9mm and 13.4%), which demonstrates the effectiveness of utilizing estimated pose as attention mask for useful feature aggregation.

4.6 Discussions

We analyze the root joint localization results on the challenging multi-person dataset MuPoTS-3D and observe several sources of large errors as shown in Figure 4: (1) Bounding boxes for two persons tend to have overlapping areas when

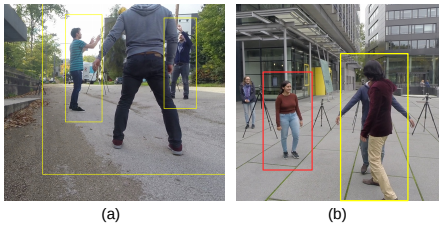

Fig. 4. Typical errors in multi-person root localization. (a) Close and overlapping bounding box regions. (b) Different sizes of target persons.

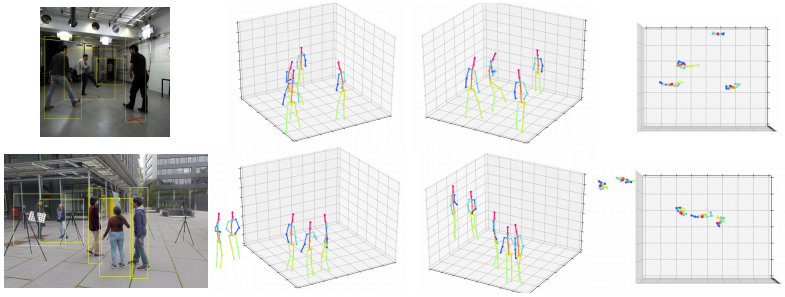

Fig. 5. Qualitative results on MuPoTS-3D dataset. Columns are: (1) image with bounding boxes (2) left-front view (3) right-front view (4) top-down view

the person closer to the camera partially occludes the other person farther away (Figure 4(a)). Masking the heatmaps with bounding box cannot effectively remove undesired regions of information and consequently the depth estimation for both persons are affected. The problem of fine-grained target person segmentation will be of interest for future research. (2) Since monocular depth estimation relies on prior knowledge such as typical scale of human bodies, estimation tends to be erroneous when the size of target person is far away from the "average" size, *e.g.*, the target is a child or a relatively short person (Figure 4(b)). Research on person 3D size estimation may complement our depth estimation task and improve the generalizability to persons of different sizes.

5 Conclusions

In this work, we proposed the Human Depth Estimation Network (HDNet), an end-to-end framework to address the problem of accurate root joint localization for multi-person 3D absolute pose estimation. Our HDNet utilizes deep features and demonstrates the capability to precisely estimate depth of root joints. We designed a human-specific pose-based feature aggregation process in the HDNet to effectively pool features from regions of human body joints. Experimental results on multiple datasets showed that our framework significantly outperforms the state-of-the-art in both root joint localization and 3D pose estimation.

References

1. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (June 2014)
2. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
3. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018)
4. Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A.: Multi-person 3d human pose estimation from monocular images. In: 3DV. pp. 405–414. IEEE (2019)
5. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: Rmpe: Regional multi-person pose estimation. In: ICCV (2017)
6. Fang, H., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: AAAI (2018)
7. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR. pp. 2002–2011 (2018)
8. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016)
10. Huang, S., Gong, M., Tao, D.: A coarse-fine network for keypoint localization. In: ICCV (2017)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 (2015)
12. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. TMAPI **36**(7), 1325–1339 (2014)
13. Iqbal, U., Molchanov, P., Breuel Juergen Gall, T., Kautz, J.: Hand pose estimation via latent 2.5 d heatmap regression. In: ECCV. pp. 118–134 (2018)
14. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
15. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: CVPR. pp. 2117–2125 (2017)
16. 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)
17. Luvizon, D.C., Picard, D., Tabia, H.: 2d/3d pose estimation and action recognition using multitask deep learning. In: CVPR. pp. 5137–5146 (2018)
18. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV. vol. 1, p. 5. IEEE (2017)
19. 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. IEEE (2017)
20. 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. IEEE (2018)
21. Moon, G., Chang, J., Lee, K.M.: Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In: ICCV (2019)
22. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NeurIPS (2017)

23. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
24. Papandreou, G., Zhu, T., Chen, L.C., Gidaris, S., Tompson, J., Murphy, K.: Personlab: Person pose estimation and instance segmentation with a bottom-up, part-based, geometric embedding model. In: ECCV (2018)
25. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR. pp. 1263–1272. IEEE (2017)
26. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR. pp. 3433–3441 (2017)
27. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. TPAMI (2019)
28. Sun, K., Xiao, B., Liu, D., Wang, J.: Deep high-resolution representation learning for human pose estimation. In: CVPR (2019)
29. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
30. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR (2016)
31. 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. pp. 8410–8419 (2018)