SMAP: Single-Shot Multi-Person Absolute 3D Pose Estimation

Jianan Zhen[1,2,*], Qi Fang[1,*], Jiaming Sun[1,2], Wentao Liu[2],
Wei Jiang[1], Hujun Bao[1], and Xiaowei Zhou[1,†]

[1] Zhejiang University
[2] SenseTime

Abstract. Recovering multi-person 3D poses with absolute scales from a single RGB image is a challenging problem due to the inherent depth and scale ambiguity from a single view. Addressing this ambiguity requires to aggregate various cues over the entire image, such as body sizes, scene layouts, and inter-person relationships. However, most previous methods adopt a top-down scheme that first performs 2D pose detection and then regresses the 3D pose and scale for each detected person individually, ignoring global contextual cues. In this paper, we propose a novel system that first regresses a set of 2.5D representations of body parts and then reconstructs the 3D absolute poses based on these 2.5D representations with a depth-aware part association algorithm. Such a single-shot bottom-up scheme allows the system to better learn and reason about the inter-person depth relationship, improving both 3D and 2D pose estimation. The experiments demonstrate that the proposed approach achieves the state-of-the-art performance on the CMU Panoptic and MuPoTS-3D datasets and is applicable to in-the-wild videos.

Keywords: Human pose estimation · 3D from a single image

1 Introduction

Recent years have witnessed an increasing trend of research on monocular 3D human pose estimation because of its wide applications in augmented reality, human-computer interaction, and video analysis. This paper aims to address the problem of estimating absolute 3D poses of multiple people simultaneously from a single RGB image. Compared to the single-person 3D pose estimation problem that focuses on recovering the root-relative pose, i.e., the 3D locations of human-body keypoints relative to the root of the skeleton, the task addressed here additionally needs to recover the 3D translation of each person in the camera coordinate system.

While there has been remarkable progress in recovering the root-relative 3D pose of a single person from an image [36, 18, 12, 8], it was not until recently

---

[*] Equal contribution. [†] Corresponding author, affiliated with the State Key Lab of CAD&CG, Zhejiang University.
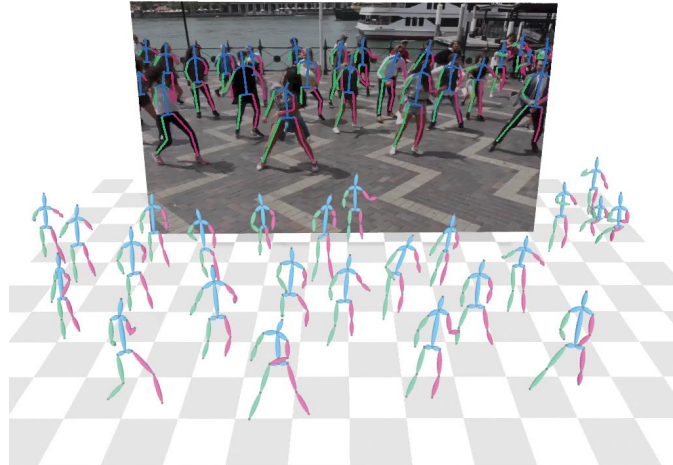
Fig. 1. We propose a novel framework named SMAP to estimate absolute 3D poses of multiple people from a single RGB image. The figure visualizes the result of SMAP on an in-the-wild image. The proposed single-shot and bottom-up design allows SMAP to leverage the entire image to infer the absolute locations of multiple people consistently, especially in terms of the ordinal depth relations.

that more attention was paid to the multi-person case. Most existing methods for multi-person 3D pose estimation extend the single-person approach with a separate stage to recover the absolute position of each detected person separately. They either use another neural network to regress the 3D translation of the person from the cropped image [22] or compute it based on the prior about the body size [6, 41, 42], which ignore the global context of the whole image. Another line of work tries to recover body positions with a ground plane constraint [19], but this approach assumes that the feet are visible, which is not always true, and accurate estimation of ground plane geometry from a single image is still an open problem.

We argue that the robust estimation of global positions of human bodies requires to aggregate the depth-related cues over the whole image, such as the 2D sizes of human bodies, the occlusion between them, and the layout of the scene. Recent advances in monocular depth estimation have shown that convolutional neural networks (CNNs) are able to predict the depth map from an RGB image [16, 13], which is particularly successful on human images [15]. This observation motivates us to directly learn the depths of human bodies from the input image instead of recovering them in a post-processing stage.

To this end, we propose a novel single-shot bottom-up approach to multi-person 3D pose estimation, which predicts absolute 3D positions and poses of multiple people in a single forward pass. We regress the root depths of human

bodies in the form of a novel root depth map, which only requires 3D pose annotations as supervision. We train a fully convolutional network to regress the root depth map, as well as 2D keypoint heatmaps, part affinity fields (PAFs), and part relative-depth maps that encode the relative depth between two joints of each body part. Then, the detected 2D keypoints are grouped into individuals based on PAFs using a part association algorithm, and absolute 3D poses are recovered with the root depth map and part relative-depth maps. The whole pipeline is illustrated in Fig. 2.

We also show that predicting depths of human bodies is beneficial for the part association and 2D pose estimation. We observe that many association errors occur when two human bodies overlap in the image. Knowing the depths of them allows us to reason about the occlusion between them when assigning the detected keypoints. Moreover, from the estimated depth, we can infer the spatial extent of each person in 2D and avoid linking two keypoints with an unreasonable distance. With these considerations, we propose a novel depth-aware part association algorithm and experimentally demonstrate its effectiveness.

To summarize, the contributions of this work are:

- A single-shot bottom-up framework for multi-person 3D pose estimation, which can reliably estimate absolute positions of multiple people by leveraging depth-relevant cues over the entire image.
- A depth-aware part association algorithm to reason about inter-person occlusion and bone-length constraints based on predicted body depths, which also benefits 2D pose estimation.
- The state-of-the-art performance on public benchmarks, with the generalization to in-the-wild images and the flexibility for both whole-body and half-body pose estimation. The code, demonstration videos and other supplementary material are available at https://zju3dv.github.io/SMAP.

2   Related work

Multi-person 2D pose. Existing methods for multi-person 2D pose estimation can be approximately divided into two classes. Top-down approaches detect human first and then estimate keypoints with a single person pose estimator [5, 7, 27, 39]. Bottom-up approaches localize all keypoints in the image first and then group them to individuals [3, 9, 23, 31, 10, 26, 25]. Cao et al. [3] propose OpenPose and use part affinity fields (PAFs) to represent the connection confidence between keypoints. They solve the part association problem with a greedy strategy. Newell et al. [23] propose an approach that simultaneously outputs detection and group assignments in the form of pixel-wise tags.

Single-person 3D pose. Researches on single-person 3D pose estimation from a single image have already achieved remarkable performances in recent years. One-stage approaches directly regress 3D keypoints from images and can leverage shading and occlusion information to resolve the depth ambiguity. Most of them are learning-based [1, 36, 28, 29, 35, 40, 8, 21]. Two-stage approaches estimate the

2D pose first and then lift it to the 3D pose, including learning-based[18, 30, 43], optimization-based [44, 38] and exemplar-based [4] methods, which can benefit from the reliable result of 2D pose estimation.

Multi-person 3D pose. For the task of multi-person 3D pose estimation, top-down approaches focus on how to integrate the pose estimation task with the detection framework. They crop the image first and then regress the 3D pose with a single-person 3D pose estimator. Most of them estimate the translation of each person with an optimization strategy that minimizes the reprojection error computed over sparse keypoints [6, 33, 34] or dense semantic correspondences [41]. Moon et al. [22] regard the area of 2D bounding box as a prior and adopt a neural network to learn a correction factor. In their framework, they regress the root-relative pose and the root depth separately. Informative cues for inferring the interaction between people may lose during the cropping operation. Another work [37] regresses the full depth map based on the existing depth estimation framework [16], but their 'read-out' strategy is not robust to 2D outliers. On the other hand, bottom-up approaches focus on how to represent pose annotations as several maps in a robust way [20, 2, 42, 19]. However, they either optimize the translation in a post-processing way or ignore the task of root localization. XNect [19] extends the 2D location map in [21] to 3D ones and estimates the translation with a calibrated camera and the ground plane constraint, but the feet may be invisible in crowded scenes and obtaining the extrinsic parameters of the camera is not practical in most applications. Another line of work tries to recover the SMPL model [42, 41], and their focus lies in using scene constraints and avoiding interpenetration, which is weakly related to our task. Taking these factors into consideration, a framework that both considers recovering the translation in a single forward pass and aggregating global features over the image will be helpful to this task.

Monocular depth estimation. Depth estimation from a single view suffers from inherent ambiguity. Nevertheless, several methods make remarkable advances in recent years [16, 13]. Li et al. [15] observe that Mannequin Challenge could be a good source for human depth datasets. They generate training data using multi-view stereo reconstruction and adopt a data-driven approach to recover a dense depth map, achieving good results. However, such a depth map lacks scale consistency and cannot reflect the real depth.

3   Methods

Fig. 2 presents the pipeline of our approach, which consists of a single-shot bottom-up framework named SMAP. With a single RGB image as input, SMAP outputs 2D representations including keypoint heatmaps and part affinity fields [3]. Additionally, it also regresses 2.5D representations including root depth map and part relative-depth maps, which encode depth information of human bodies. Then, a depth-aware part association algorithm is proposed to assign detected
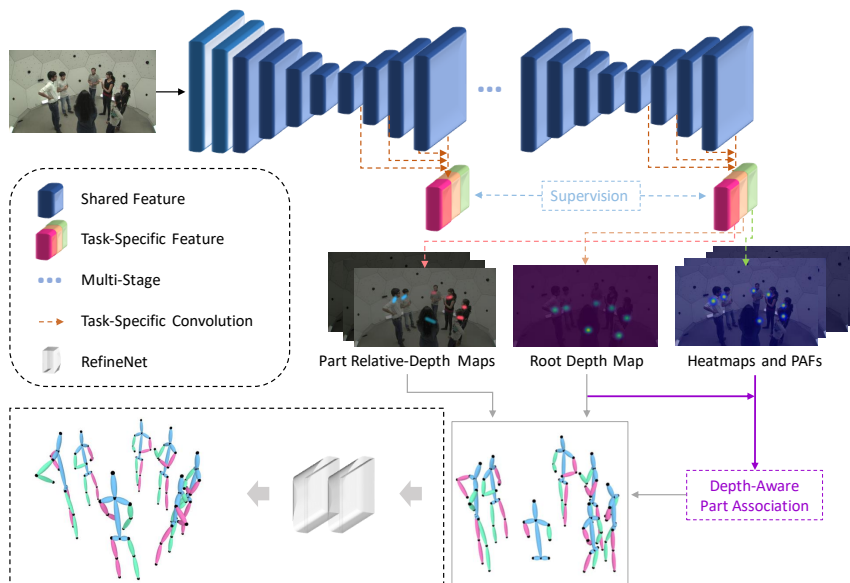
Fig. 2. Overview of the proposed approach. Given a single image, our single-shot network SMAP regresses several intermediate representations including 2D keypoint heatmaps, part affinity fields (PAFs), root depth map, and part relative-depth maps (Red means the child joint has a larger depth than its parent joint, and blue means the opposite). With a new depth-aware part association algorithm, body parts belonging to the same person are linked. With all these intermediate representations combined, absolute 3D poses of all people can be recovered. Finally, an optional RefineNet can be used to further refine the recovered 3D poses and complete invisible keypoints.

2D keypoints to individuals, depending on an ordinal prior and an adaptive bone-length constraint. Based on these results, the absolute 3D pose of each person can be reconstructed with a camera model. Individual modules of our system are introduced below.

3.1   Intermediate representations

Given the input image, SMAP regresses the following intermediate representations, based on which 3D poses will be reconstructed:

Root depth map. As the number of people in an input image is unknown, we propose to represent the absolute depths for all human bodies in the image by a novel root depth map. The root depth map has the same size as the input image. The map values at the 2D locations of root joints of skeletons equal their absolute depths. An example is shown in Fig. 2. In this way, we are able to represent the depths of multiple people without predefining the number of

people. During training, we only supervise the values of root locations. The proposed root depth map can be learned together with other cues within the same network as shown in Fig. 2 and only requires 3D poses (instead of full depth maps) as supervision, making our algorithm very efficient in terms of both model complexity and training data.

It is worth noting that visual perception of object scale and depth depends on the size of field of view (FoV), i.e., the ratio between the image size and the focal length. If two images are obtained with different FoVs, the same person at the same depth will occupy different proportions in these two images and seem to have different depths for the neural network, which may mislead the learning of depth. Thus, we normalize the root depth by the size of FoV as follows:

$$\widetilde{Z} = Z\frac{w}{f}, \tag{1}$$

where $\widetilde{Z}$ is the normalized depth, Z is the original depth, and f and w are the focal length and the image width both in pixels, respectively. So w/f is irrelevant to image resolution, but equals to the ratio between the physical size of image sensor and the focal length both in millimeters, i.e., FoV. The normalized depth values can be converted back to metric values in inference.

Keypoint heatmaps. Each keypoint heatmap indicates the probable locations of a specific type of keypoints for all people in the image. Gaussian distribution is used to model uncertainties at the corresponding location.

Part affinity fields (PAFs). PAFs proposed in [3] include a set of 2D vector fields. Each vector field corresponds to a type of body part where the vector at each pixel represents the 2D orientation of the corresponding body part.

Part relative-depth map. Besides the root depth, we also need depth values of other keypoints to reconstruct a 3D pose. Instead of predicting absolute depth values, for other keypoints we only regress their relative depths compared to their parent nodes in the skeleton, which are represented by part relative-depth maps. Similar to PAFs, each part relative-depth map corresponds to a type of body part, and every pixel that belongs to a body part encodes the relative depth between two joints of the corresponding body part. This dense representation provides rich information to reconstruct a 3D skeleton even if some keypoints are invisible.

Network architecture. We use Hourglass [24] as our backbone and modify it to a multi-task structure with multiple branches that simultaneously output the above representations as illustrated in Fig. 2. Suppose the number of predefined joints is J. Then, there are $4J - 2$ channels in total (heatmaps and PAFs: $J + 2(J - 1)$, root depth map: 1, part relative-depth map: $J - 1$). Each output branch in our network only consists of two convolutional layers. Inspired by [14], we adopt multi-scale intermediate supervision. The L_1 loss is used to supervise the root depths and L_2 losses on other outputs. The effect of the network size and the multi-scale supervision will be validated in the experiments.
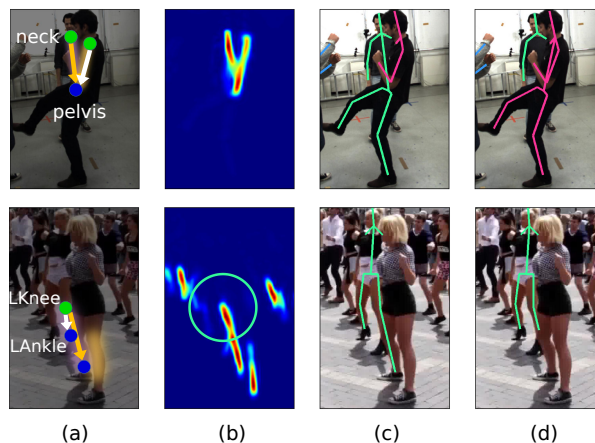
Fig. 3. Depth-aware part association. From left to right: candidate links, part affinity fields, pose estimation results from [3], and our results with depth-aware part association. The example in the first row shows the effect of ordinal prior. Under this circumstance, [3] assigns the pelvis to the occluded person while we give priority to the front person. The example in the second row shows the effect of the adaptive bone-length threshold, indicated by the green circle. As a noisy response occurs at the right ankle of the person it doesn't belong to, [3] will induce a false connection while our algorithm will not.

3.2   Depth-aware part association

Given 2D coordinates of keypoints from keypoint heatmaps after non-maximum suppression, we need to associate detected joints with corresponding individuals. Cao et al. [3] propose to link joints greedily based on the association scores given by PAFs. Basically, we follow their method to calculate association scores. However, PAFs scores might be unreliable due to occlusion between people. Fig. 3 shows two examples where the above association strategy fails.

We propose to leverage the estimated depth maps to address the part association ambiguities caused by inter-person occlusion.

Ordinal prior. A key insight to solve the occlusion issue is to give priority to the unoccluded person when assigning joints. The occlusion status can be inferred from the depth map. Therefore, we sort root joints from near to far according to the predicted root depth, instead of following the order of PAFs scores. Our association process starts with the root and proceeds along the skeleton tree successively.

Adaptive bone-length constraint. To avoid linking two keypoints with an unreasonable distance, Cao et al. [3] constrain the association with a threshold defined as half of the image size. However, such a fixed threshold is ineffective as the 2D bone length depends on its depth. We adopt an adaptive bone-length

threshold to penalize the unreasonable association, depending on predicted body depths. For each part, we compute the mean bone length in the training set in advance. Then, its maximal length in 2D is computed and used as the distance threshold for the corresponding part:

$$d_{cons} = \lambda \cdot \frac{D_{bone} \cdot f}{Z \cdot w} = \lambda \cdot \frac{D_{bone}}{\widetilde{Z}} \ , \tag{2}$$

where D_{bone} is the 3D average length of a limb, $\widetilde{Z}$ is the normalized root depth predicted by our network, and λ is a relaxation factor. d_{cons} is used to filter unreasonable links. From Eq. 2 we can see that the adaptive threshold is not affected by camera intrinsic parameters or image resizing and is only related to the depth value estimated by our network and the statistical bone length.

As the association can benefit from depth information, we call it depth-aware part association. Fig. 3(d) shows our qualitative results. The proposed scheme will also be validated by experiments.

3.3   3D pose reconstruction

Reconstruction. Following the connection relations obtained from part association, relative depths from child nodes to parent nodes can be read from the corresponding locations of part relative-depth maps. With the root depth and relative depths of body parts, we are able to compute the depth of each joint. Given 2D coordinates and joint depths, the 3D pose can be recovered through the perspective camera model:

$$\left[X, Y, Z\right]^T = Z K^{-1} \left[x, y, 1\right]^T \ , \tag{3}$$

where $[X, Y, Z]^T$ and $[x, y]^T$ are 3D and 2D coordinates of a joint respectively. K is the camera intrinsic matrix, which is available in most applications, e.g., from device specifications. In our experiments, we use the focal lengths provided by the datasets (same as [22]). For internet images with unknown focal lengths, we use a default value which equals the input image width in pixels. Note that the focal length will not affect the predicted ordinal depth relations between people.

Refinement. The above reconstruction procedure may introduce two types of errors. One is the cumulative error in the process of joint localization due to the hierarchical skeleton structure, and the other is caused by back projection when the depth is not accurate enough to calculate X and Y coordinates of 3D pose. Moreover, severe occlusion and truncation frequently occur in crowded scenes, which make some keypoints invisible. Therefore, we use an additional neural network named RefineNet to refine visible keypoints and complete invisible keypoints for each 3D pose. RefineNet consists of five fully connected layers. The inputs are 2D pose and 3D root-relative pose while the output is the refined 3D root-relative pose. The coordinates of invisible keypoints in the input are set to be zero. Note that RefineNet doesn't change the root depths.

4   Experiments

We evaluate the proposed approach on two widely-used datasets and compare it to previous approaches. Besides, we provide thorough ablation analysis to validate our designs.

4.1   Datasets

CMU Panoptic   [11] is a large-scale dataset that contains various indoor social activities, captured by multiple cameras. Mutual occlusion between individuals and truncation makes it challenging to recover 3D poses. Following [41], we choose two cameras (16 and 30), 9600 images from four activities (Haggling, Mafia, Ultimatum, Pizza) as our test set, and 160k images from different sequences as our training set.

MuCo-3DHP and MuPoTS-3D [20]. MuCo-3DHP is an indoor multi-person dataset for training, which is composited from single-person datasets. MuPoTS-3D is a test set consisting of indoor and outdoor scenes with various camera poses, making it a convincing benchmark to test the generalization ability.

4.2   Implementation details

We adopt Adam as optimizer with 2e-4 learning rate, and train two models for 20 epochs on the CMU Panoptic and MuCo-3DHP datasets separately, mixed with COCO data [17]. The batch size is 32 and 50% data in each mini-batch is from COCO (same as [20, 22]). Images are resized to a fixed size $832{\times}512$ as the input to the network. Note that resizing doesn't change FoV. Since the COCO dataset lacks 3D pose annotations, weights of 3D losses are set to zero when the COCO data is fed.

4.3   Evaluation metrics

MPJPE. MPJPE measures the accuracy of the 3D root-relative pose. It calculates the Euclidean distance between the predicted and the groundtruth joint locations averaged over all joints.

RtError. Root Error (RtError) is defined as the Euclidean distance between the predicted and the groundtruth root locations.

3DPCK. 3DPCK is the percentage of correct keypoints. A keypoint is declared correct if the Euclidean distance between predicted and groundtruth coordinates is smaller than a threshold (15cm in our experiments). PCK_{rel} measures relative pose accuracy with root alignment; PCK_{abs} measures absolute pose accuracy without root alignment; and PCK_{root} only measures the accuracy of root joints. AUC means the area under curve of 3DPCK over various thresholds.

Table 1. Results on the Panoptic dataset. For [22], we used the code provided by the authors and trained it on the Panoptic dataset. *The average of [42] is recalculated following the standard practice in [41], i.e., average over activities.

	Method	Haggling	Mafia	Ultim.	Pizza	Average
MPJPE	PoPa et al. [32]	217.9	187.3	193.6	221.3	203.4
	Zanfir et al. [41]	140.0	165.9	150.7	156.0	153.4
	Moon et al. [22]	89.6	91.3	79.6	90.1	87.6
	Zanfir et al. [42]	72.4	78.8	66.8	94.3	78.1*
	Ours w/o Refine	71.8	72.5	65.9	82.1	73.1
	Ours	**63.1**	**60.3**	**56.6**	**67.1**	**61.8**
RtError	Zanfir et al. [41]	257.8	409.5	301.1	294.0	315.5
	Moon et al. [22]	160.2	151.9	177.5	127.7	154.3
	Ours	**84.7**	**87.7**	**91.2**	**78.5**	**85.5**
PCOD	Moon et al. [22]	92.3	93.7	95.2	94.2	93.9
	Ours	**97.8**	**98.5**	**97.6**	**99.6**	**98.4**

Table 2. Results on the MuPoTS-3D dataset. All numbers are average values over 20 activities.

		Matched people					All people	
	Method	PCK_{rel}	PCK_{abs}	PCK_{root}	AUC_{rel}	PCOD	PCK_{rel}	PCK_{abs}
top down	Rogez. [33]	62.4	-	-	-	-	53.8	-
	Rogez. [34]	74.0	-	-	-	-	70.6	-
	Dabral. [6]	74.2	-	-	-	-	71.3	-
	Moon. [22]	**82.5**	31.8	31.0	40.9	92.6	**81.8**	31.5
bottom up	Mehta. [20]	69.8	-	-	-	-	65.0	-
	Mehta. [19]	75.8	-	-	-	-	70.4	-
	Ours	**80.5**	**38.7**	**45.5**	**42.7**	**97.0**	**73.5**	**35.4**

PCOD. We propose a new metric named the percentage of correct ordinal depth (PCOD) relations between people. The insight is that predicting absolute depth from a single view is inherently ill-posed, while consistent ordinal relations between people are more meaningful and suffice many applications. For a pair of people (i, j), we compare their root depths and divide the ordinal depth relation into three classes: closer, farther, and roughly the same (within 30cm). PCOD equals the classification accuracy of predicted ordinal depth relations.

4.4 Comparison with state-of-the-art methods

CMU Panoptic. Table 1 demonstrates quantitative comparison between state-of-the-art methods and our model. It indicates that our model outperforms previous methods in all metrics by a large margin. In particular, the error on the Pizza sequence decreases significantly compared with the previous work. As the Pizza sequence shares no similarity with the training set, this improvement shows our generalization ability.
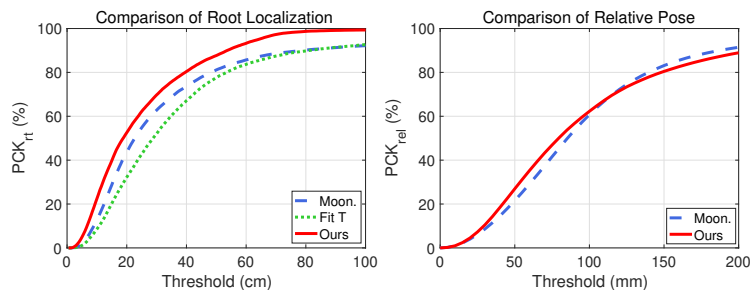
Fig. 4. Comparisons of root localization and relative pose. The curves are PCK_{root} and PCK_{rel} over different thresholds on the MuPoTS-3D dataset. Blue: result of [22]. Green: estimating the translation by minimizing the reprojection error. Red: our result.
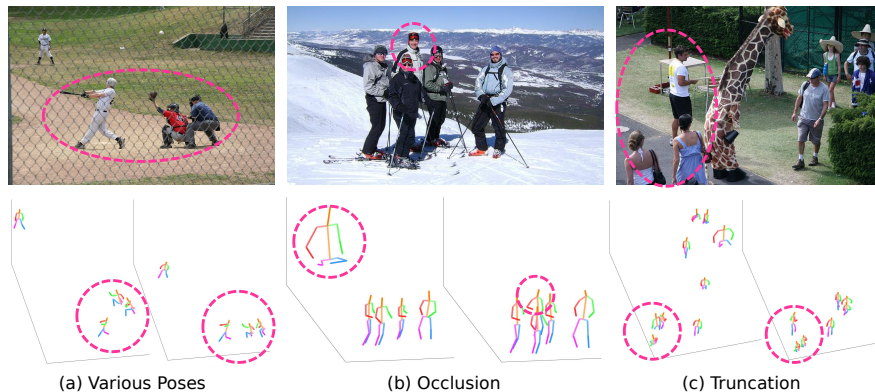

(a) Various Poses   (b) Occlusion   (c) Truncation

Fig. 5. Qualitative comparison. The results of three example images. For each example, the top row shows the input image, and the bottom row shows the results of [22] (left) and the proposed method (right), respectively. The red circles highlight the difference in localization of human bodies between two methods.

MuPoTS-3D. We follow the protocol of [22]. Additionally, PCK_{root} and PCOD are used to evaluate the 3D localization of people. In terms of the absolute pose which we are more concerned with, it can be observed from Table 2 that our model is superior to [22] in relevant metrics including PCK_{abs}, PCK_{root} and PCOD by a large margin. It also demonstrates that our model has higher PCK_{rel} compared with all bottom-up methods and most top-down methods except [22]. Note that we achieve higher AUC_{rel} compared to [22] for the relative 3D pose of matched people.

Comparison with top-down methods. We provide additional analysis to compare our single-shot bottom-up method to the state-of-the-art top-down method [22]. Fig. 4 shows thorough comparisons in terms of PCK_{root} and PCK_{rel}.

For root localization, we compare to two methods: 1) regressing the scale from each cropped bounding box using a neural network as in [22] and 2) estimating the 3D translation by optimizing reprojection error with the groundtruth 2D pose and the estimated relative 3D pose from [22] ('FitT' in Fig. 4). We achieve better PCK_{root} over various thresholds than both of them. Notably, we achieve roughly 100% accuracy with a threshold 1m. As for relative pose estimation, [22] achieves higher PCK_{rel} (@15cm) as it adopts a separate off-the-shelf network [36] that is particularly optimized for relative 3D pose estimation. Despite that, we obtain better PCK_{rel} when the threshold is smaller and higher AUC_{rel}.

Fig. 5 shows several scenarios (various poses, occlusion, and truncation) in which the top-down method [22] may fail as it predicts the scale for each detected person separately and ignore global context. Instead, the proposed bottom-up design is able to leverage features over the entire image instead of only using cropped features in individual bounding boxes.

Furthermore, our running time and memory remain almost unchanged with the number of people in the image while those of [22] grow faster with the number of people due to its top-down design, as shown in the supplementary material.

Depth estimation. Apart from our method, there are two alternatives for depth estimation: 1) regressing the full depth map rather than the root depth map. 2) using the cropped image as the input to the network rather than the whole image. For the first alternative, since there is no depth map annotation in existing multi-person outdoor datasets, we use the released model of the state-of-the-art human depth estimator [15], which is particularly optimized for human depth estimation trained on a massive amount of in-the-wild 'frozen people' videos. For the second alternative, [22] is the state-of-the-art method that estimates root depth from the cropped image, so we compare with it. Fig. 6 demonstrates scatter plots of the groundtruth root depth versus the predicted root depth of three methods on the MuPoTS-3D dataset. Ideally, the estimated depths should be linearly correlated to the ground truth, resulting a straight line in the scatter plot. Our model shows better consistency than baselines. Note that, while the compared methods are trained on different datasets, the images in the test set MuPoTS-3D are very different from the training images for all methods. Though not rigorous, this comparison is still reasonable to indicate the performance of these methods when applied to unseen images.

4.5   Ablation analysis

Architecture. Table 3 shows how different designs of our framework affect the multi-person 3D pose estimation accuracy: 1) the performance of our model will degrade severely without depth normalization. As we discussed in Section 3.1, normalizing depth values by the size of FoV makes depth learning easier. 2) Multi-scale supervision is beneficial. 3) To show that our performance gain in terms of the absolute 3D pose is mostly attributed to our single-shot bottom-up design rather than the network size, we test with a smaller backbone. The results
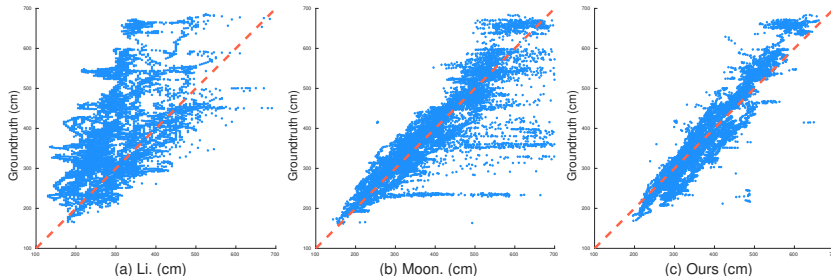
(a) Li. (cm)   (b) Moon. (cm)   (c) Ours (cm)

Fig. 6. Comparison with alternative depth estimation methods. The scatter plots show the consistency of root depth estimation on the MuPoTS-3D dataset. The X and Y axes are the predicted, groundtruth root depth, respectively. The dashed line means the ideal result, i.e., estimation equals ground truth. (a) 'Read-out' root depths from the full depth map estimated by [15]. (b) State-of-the-art top-down approach [22]. (c) Our approach.

Table 3. Ablation study of the structure design on the MuPoTS-3D dataset. The default backbone is Hourglass model with three stages, and 'Smaller Backbone' means one-stage model.

Design	Recall	PCK_{root}	PCK_{abs}	PCK_{rel}	PCOD
Full Model	**92.3**	**45.5**	**38.7**	**80.5**	**97.0**
No Normalization	92.3	5.7	8.7	78.9	95.7
No Multi-scale Supervision	92.1	45.2	36.2	75.4	93.1
No RefineNet	92.3	45.5	34.7	70.9	97.0
Smaller Backbone	91.1	43.8	35.1	75.7	96.4

show that, even with a one-stage hourglass network, our method still achieves higher PCK_{root} and PCK_{abs} than the top-down method [22].

Part association. To compare the proposed depth-aware part association with the 2D part association in [3], we evaluate relevant metrics on Panoptic and MuPoTS-3D datasets. Note that the threshold of 2DPCK is the half of the head size. Table 4 lists the results (2DPA vs. DAPA) and reveals that our depth-aware part association outperforms the 2D part association in all these metrics. Besides, Fig. 3 shows some qualitative examples.

RefineNet. Table 1 and 3 show that RefineNet is able to improve both relative and absolute pose estimation. It is able to complete invisible keypoints and refine visible keypoints with a learned 3D pose prior. The improvement is more significant on the Panoptic dataset since the training and test images are captured by cameras with similar views.

Please refer to the supplementary material for more experimental details and results.

Table 4. Ablation study of the part association. '2DPA' means the 2D part association proposed by [3]. 'DAPA' means the depth-aware part association we proposed. Both of them are based on the same heatmaps and PAFs results.

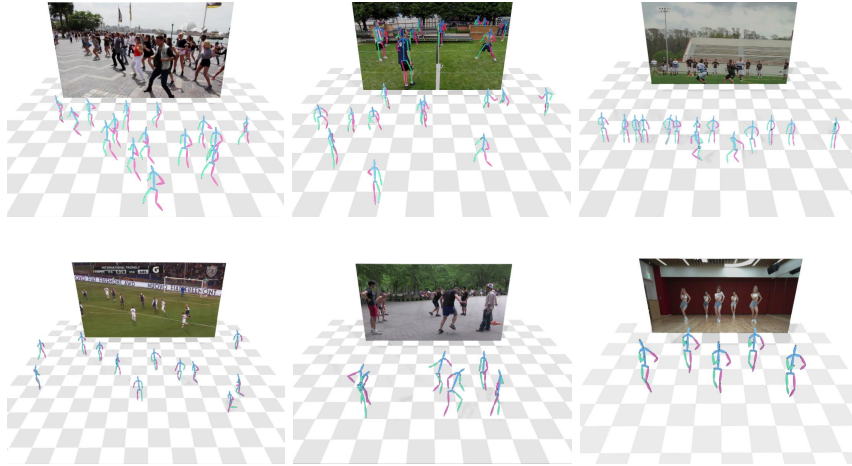|	Panoptic		MuPoTS-3D				
	Recall	2DPCK	Recall	PCK_{root}	PCK_{abs}	PCK_{rel}	PCOD
2DPA	94.3	92.4	92.1	45.3	38.6	80.2	96.5
DAPA	**96.4**	**93.1**	**92.3**	**45.5**	**38.7**	**80.5**	**97.0**

Fig. 7. Qualitative results on in-the-wild images from the Internet.

5   Conclusion

We proposed a novel single-shot bottom-up framework to estimate absolute multi-person 3D poses from a single RGB image. The proposed framework uses a fully convolutional network to regress a set of 2.5D representations for multiple people, from which the absolute 3D poses can be reconstructed. Additionally, benefited from the depth estimation of human bodies, a novel depth-aware part association algorithm was proposed and proven to benefit 2D pose estimation in crowd scenes. Experiments demonstrated state-of-the-art performance as well as generalization ability of the proposed approach.

References

1. Alp Güler, R., Neverova, N., Kokkinos, I.: Densepose: Dense human pose estimation in the wild. In: CVPR (2018)
2. Benzine, A., Luvison, B., Pham, Q.C., Achard, C.: Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In: ICIP (2019)
3. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
4. Chen, C.H., Ramanan, D.: 3d human pose estimation= 2d pose estimation+ matching. In: CVPR (2017)
5. Chen, Y., Wang, Z., Peng, Y., Zhang, Z., Yu, G., Sun, J.: Cascaded pyramid network for multi-person pose estimation. In: CVPR (2018)
6. Dabral, R., Gundavarapu, N.B., Mitra, R., Sharma, A., Ramakrishnan, G., Jain, A.: Multi-person 3d human pose estimation from monocular images. In: 3DV (2019)
7. Fang, H.S., Xie, S., Tai, Y.W., Lu, C.: RMPE: Regional multi-person pose estimation. In: ICCV (2017)
8. Guler, R.A., Kokkinos, I.: Holopose: Holistic 3d human reconstruction in-the-wild. In: CVPR (2019)
9. Hidalgo, G., Raaj, Y., Idrees, H., Xiang, D., Joo, H., Simon, T., Sheikh, Y.: Single-network whole-body pose estimation. In: ICCV (2019)
10. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., Schiele, B.: Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In: ECCV (2016)
11. Joo, H., Simon, T., Li, X., Liu, H., Tan, L., Gui, L., Banerjee, S., Godisart, T.S., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social interaction capture. TPAMI (2017)
12. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: CVPR (2018)
13. Lee, J.H., Kim, C.S.: Monocular depth estimation using relative depth maps. In: CVPR (2019)
14. Li, W., Wang, Z., Yin, B., Peng, Q., Du, Y., Xiao, T., Yu, G., Lu, H., Wei, Y., Sun, J.: Rethinking on multi-stage networks for human pose estimation. arXiv preprint arXiv:1901.00148 (2019)
15. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: CVPR (2019)
16. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: CVPR (2018)
17. 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)
18. Martinez, J., Hossain, R., Romero, J., Little, J.J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
19. 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. TOG (2020)
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 (2018)
21. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. TOG (2017)

22. 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)
23. Newell, A., Huang, Z., Deng, J.: Associative embedding: End-to-end learning for joint detection and grouping. In: NeurIPS (2017)
24. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECCV (2016)
25. Nie, X., Zhang, J., Yan, S., Feng, J.: Single-stage multi-person pose machines. In: ICCV (2019)
26. 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)
27. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: CVPR (2017)
28. Pavlakos, G., Zhou, X., Daniilidis, K.: Ordinal depth supervision for 3d human pose estimation. In: CVPR (2018)
29. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR (2017)
30. Pavllo, D., Feichtenhofer, C., Grangier, D., Auli, M.: 3d human pose estimation in video with temporal convolutions and semi-supervised training. In: CVPR (2019)
31. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: CVPR (2016)
32. Popa, A.I., Zanfir, M., Sminchisescu, C.: Deep multitask architecture for integrated 2d and 3d human sensing. In: CVPR (2017)
33. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: CVPR (2017)
34. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net++: Multi-person 2d and 3d pose detection in natural images. TPAMI (2019)
35. Sun, X., Shang, J., Liang, S., Wei, Y.: Compositional human pose regression. In: ICCV (2017)
36. Sun, X., Xiao, B., Wei, F., Liang, S., Wei, Y.: Integral human pose regression. In: ECCV (2018)
37. Véges, M., Lőrincz, A.: Absolute human pose estimation with depth prediction network. In: IJCNN (2019)
38. Xiang, D., Joo, H., Sheikh, Y.: Monocular total capture: Posing face, body, and hands in the wild. In: CVPR (2019)
39. Xiao, B., Wu, H., Wei, Y.: Simple baselines for human pose estimation and tracking. In: ECCV (2018)
40. 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)
41. 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)
42. 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)
43. Zhao, L., Peng, X., Tian, Y., Kapadia, M., Metaxas, D.N.: Semantic graph convolutional networks for 3d human pose regression. In: CVPR (2019)
44. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: CVPR (2016)