Monocular 3D Multi-Person Pose Estimation by Integrating Top-Down and Bottom-Up Networks

Yu Cheng[1], Bo Wang[2], Bo Yang[2], Robby T. Tan[1,3]
[1]National University of Singapore
[2]Tencent Game AI Research Center
[3]Yale-NUS College
e0321276@u.nus.edu, {bohawkwang,brandonyang}@tencent.com, robby.tan@nus.edu.sg

Abstract

In monocular video 3D multi-person pose estimation, inter-person occlusion and close interactions can cause human detection to be erroneous and human-joints grouping to be unreliable. Existing top-down methods rely on human detection and thus suffer from these problems. Existing bottom-up methods do not use human detection, but they process all persons at once at the same scale, causing them to be sensitive to multiple-persons scale variations. To address these challenges, we propose the integration of top-down and bottom-up approaches to exploit their strengths. Our top-down network estimates human joints from all persons instead of one in an image patch, making it robust to possible erroneous bounding boxes. Our bottom-up network incorporates human-detection based normalized heatmaps, allowing the network to be more robust in handling scale variations. Finally, the estimated 3D poses from the top-down and bottom-up networks are fed into our integration network for final 3D poses. Besides the integration of top-down and bottom-up networks, unlike existing pose discriminators that are designed solely for a single person, and consequently cannot assess natural interperson interactions, we propose a two-person pose discriminator that enforces natural two-person interactions. Lastly, we also apply a semi-supervised method to overcome the 3D ground-truth data scarcity. Quantitative and qualitative evaluations show the effectiveness of the proposed method. Our code is available publicly. [1]

1. Introduction

Estimating 3D multi-person poses from a monocular video has drawn increasing attention due to its importance for real-world applications (e.g., [32, 28, 2, 7]). Unfortunately, it is generally still challenging and an open problem, particularly when multiple persons are present in the scene. Multiple persons can generate inter-person occlusion, which causes human detection to be erroneous. Moreover, multiple persons in a scene are likely in close contact with each other and interact, which makes human-joints grouping unreliable.

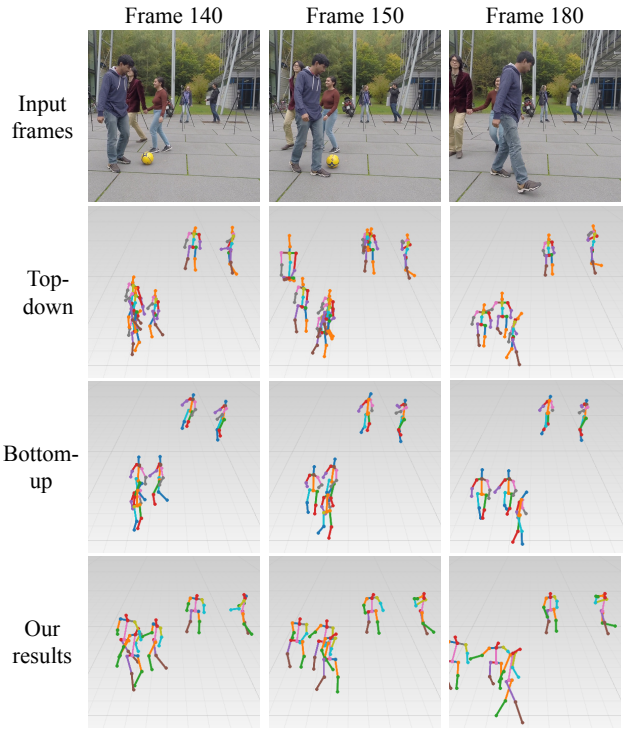

Figure 1. Incorrect 3D multi-person pose estimation from existing top-down (2nd row) and bottom-up (3rd row) methods. The top-down method is RootNet [32], the bottom-up method is SMAP [53]. The input images are from MuPoTS-3D dataset [29]. The top-down method suffers from inter-person occlusion and the bottom-up method is sensitive to scale variations (i.e., the 3D poses of the two persons in the back are inaccurately estimated). Our method substantially outperforms the state-of-the-art.

[1] https://github.com/3dpose/3D-Multi-Person-Pose

Although existing 3D human pose estimation methods (e.g., [31, 52, 35, 13, 36, 8, 7]) show promising results on single-person datasets like Human3.6M [16] and HumanEva [40], these methods do not perform well in 3D multi-person scenarios. Generally, we can divide existing methods into two approaches: top-down and bottom-up. Existing top-down 3D pose estimation methods rely considerably on human detection to localize each person, prior to estimating the joints within the detected bounding boxes, e.g., [36, 8, 32]. These methods show promising performance for single-person 3D-pose estimation [36, 8], yet since they treat each person individually, they have no awareness of non-target persons and the possible interactions. When multiple persons occlude each other, human detection also become unreliable. Moreover, when target persons are closely interacting with each other, the pose estimator may be misled by the nearby persons, e.g., predicted joints may come from the nearby non-target persons.

Recent bottom-up methods (e.g., [53, 24, 22]) do not use any human detection and thus can produce results with higher accuracy when multiple persons interact with each other. These methods consider multiple persons simultaneously and, in many cases, better distinguish the joints of different persons. Unfortunately, without using detection, bottom-up methods suffer from the scale variations, and the pose estimation accuracy is compromised, rendering inferior performance compared with top-down approaches [5]. As shown in Figure 1, neither top-down nor bottom-up approach alone can handle all the challenges at once, particularly the challenges of: inter-person occlusion, close interactions, and human-scale variations. Therefore, in this paper, our goal is to integrate the top-down and bottom-up approaches to achieve more accurate and robust 3D multi-person pose estimation from a monocular video.

To achieve this goal, we introduce a top-down network to estimate human joints inside each detected bounding box. Unlike existing top-down methods that only estimate one human pose given a bounding box, our top-down network predicts 3D poses for all persons inside the bounding box. The joint heatmaps from our top-down network is feed to our bottom-up network, so that our bottom network can be more robust in handling the scale variations. Finally, we feed the estimated 3D poses from both top-down and bottom-up networks into our integration network to obtain the final estimated 3D poses given an image sequence.

Moreover, unlike existing methods' pose discriminators, which are designed solely for single person, and consequently cannot enforce natural inter-person interactions, we propose a two-person pose discriminator that enforces two-person natural interactions. Lastly, semi-supervised learning is used to mitigate the data scarcity problem where 3D ground-truth data is limited.

In summary, our contributions are listed as follows.

- We introduce a novel two-branch framework, where the top-down branch detects multiple persons and the bottom-up branch incorporates the normalized image patches in its process. Our framework gains benefits from the two branches, and at the same time, overcomes their shortcomings.

- We employ multi-person pose estimation for our top-down network, which can effectively handle the inter-person occlusion and interactions caused by detection errors.

- We incorporate human detection information into our bottom-up branch so that it can better handle the scale variation, which addresses the problem in existing bottom-up methods.

- Unlike the existing discriminators that focus on single person pose, we introduce a novel discriminator that enforces the validity of human poses of close pairwise interactions in the camera-centric coordinates.

2. Related Works

Top-Down Monocular 3D Human Pose Estimation Existing top-down 3D human pose estimation methods commonly use human detection as an essential part of their methods to estimate person-centric 3D human poses [27, 34, 31, 36, 8, 9, 7]. They demonstrate promising performance on single-person evaluation datasets [16, 40], unfortunately the performance decreases in multi-person scenarios, due to inter-person occlusion or close interactions [31, 8]. Moreover, the produced person-centric 3D poses cannot be used for multi-person scenarios, where camera-centric 3D-pose estimation is needed. Top-down methods process each person independently, leading to inadequate awareness of the existence of other persons nearby. As a result, they perform poorly on multi-person videos where inter-person occlusion and close interactions are commonly present. Rogez et al. [38, 39] develop a pose proposal network to generate bounding boxes and then perform pose estimation individually for each person. Recently, unlike previous methods that perform person-centric pose estimation, Moon et al. [32] propose a top-down 3D multi-person pose-estimation method that can estimate the poses for all persons in an image in the camera-centric coordinates. However, the method still relies on detection and process each person independently; hence it is likely to suffer from inter-person occlusion and close interactions.

Bottom-Up Monocular 3D Human Pose Estimation A few bottom-up methods have been proposed [10, 53, 28, 22, 24]. Fabbri et al. [10] introduce an encoder-decoder framework to compress a heatmap first, and then decompress it back to the original representations in the test time for fast

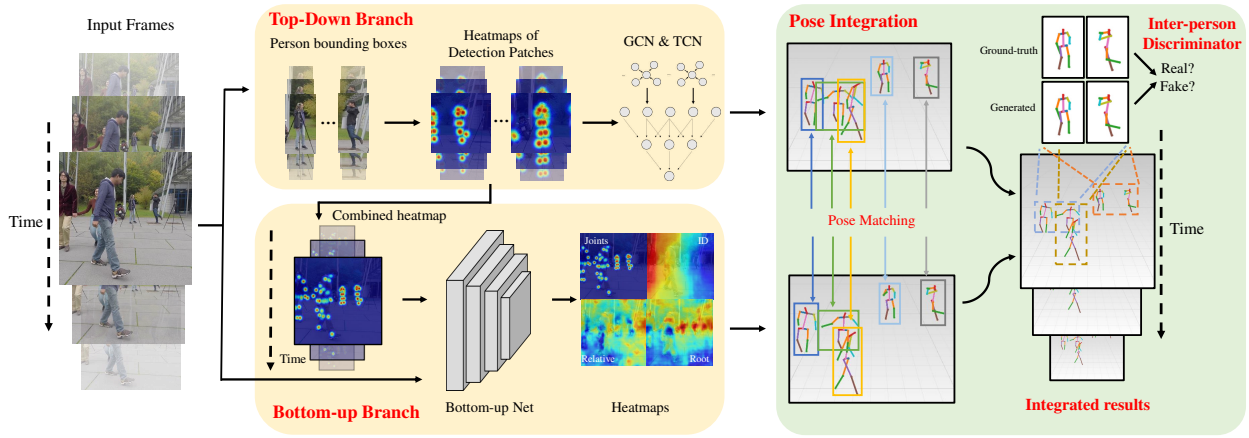

Figure 2. The overview of our framework. Our proposed method comprises three components: 1) A top-down branch to estimate fine-grained instance-wise 3D pose. 2) A bottom-up branch to generate global-aware camera-centric 3D pose. 3) An integration network to generate final estimation based on paired poses from top-down and bottom-up to take benefits from both branches. Note that the semi-supervised learning part is a training strategy so it is not included in this figure.

HD image processing. Mehta et al. [28] propose to identify individual joints, compose full-body joints, and enforce temporal and kinematic constraints in three stages for real-time 3D motion capture. Li et al. [22] develop an integrated method with lower computation complexity for human detection, person-centric pose estimation, and human depth estimation from an input image. Lin et al. [24] formulate the human depth regression as a bin index estimation problem for multi-person localization in the camera coordinate system. Zhen et al. [53] estimate the 2.5D representation of body parts first and then reconstruct camera-centric multi-person 3D poses. These methods benefit from the nature of the bottom-up approach, which can process multiple persons simultaneously without relying on human detection. However, since all persons are processed at the same scale, these methods are inevitably sensitive to human scale variations, which limits their applicability on wild videos.

Top-Down and Bottom-Up Combination Earlier non-deep learning methods exploring the combination of top-down and bottom-up approaches for human pose estimation are in the forms of data-driven belief propagation, different classifiers for joint location and skeleton, or probabilistic Gaussian mixture modelling [15, 48, 21]. Recent deep learning based methods that attempt to make use of both top-down and bottom-up information are mainly on estimating 2D poses [14, 43, 3, 23]. Hu and Ramanan [14] propose a hierarchical rectified Gaussian model to incorporate top-down feedback with bottom-up CNNs. Tang et al. [43] develop a framework with bottom-up inference followed by top-down refinement based on a compositional model of the human body. Cai et al. [3] introduce a spatial-temporal graph convolutional network (GCN) that uses both bottom-up and top-down features. These methods explore to benefit from top-down and bottom-up information. However, they are not suitable for 3D multi-person pose estimation because the fundamental weaknesses in both top-down and bottom-up methods are not addressed completely, which include inter-person occlusion caused detection and joints grouping errors, and the scale variation issue. Li et al. [23] adopt LSTM and combine bottom-up heatmaps with human detection for 2D multi-person pose estimation. They address occlusion and detection shift problems. Unfortunately, they use a bottom-up network and only add the detection bounding box as the top-down information to group the joints. Hence, their method is essentially still bottom-up and thus still vulnerable to human scale variations.

3. Proposed Method

Fig. 2 shows our pipeline, which consists of three major parts to accomplish the multi-person camera-centric 3D human pose estimation: a top-down network for fine-grained instance-wise pose estimation, a bottom-up network for global-aware pose estimation, and an integration network to integrate the estimations of top-down and bottom-up branches with inter-person pose discriminator. Moreover, a semi-supervised training process is proposed to enhance the 3D pose estimation based on reprojection consistency.

3.1. Top-Down Network

Given a human detection bounding box, existing top-down methods estimate full-body joints of one person. Consequently, if there are multiple persons inside the box or partially out-of-bounding box body parts, the full-body joint estimation are likely to be erroneous. Figure 3 shows such failure examples of existing methods. In contrast, our method produces the heatmaps for all joints inside the

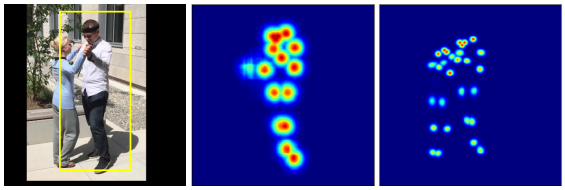

Figure 3. Examples of estimated heatmaps of human joints. The left image shows the input frame overlaid with inaccurate detection bounding box (i.e., only one person detected). The middle image shows the estimated heatmap of existing top-down methods. The right image shows the heatmap of our top-down branch.

bounding box (i.e., enlarged to accommodate inaccurate detection), and estimate the ID for each joint to group them into corresponding persons, similar to [33].

Given an input video, for every frame we apply a human detector [12], and crop the image patches based on the detected bounding boxes. A 2D pose detector [5] is applied to each patch to generate heatmaps for all human joints, such as shoulder, pelvis, ankle, and etc. Specifically, our top-down loss of 2D pose heatmap is an L2 loss between the predicted and ground-truth heatmaps, formulated as:

$$L_{hmap}^{TD} = |H - \tilde{H}|_2^2, \tag{1}$$

where H and $\tilde{H}$ are the predicted and ground-truth heatmaps, respectively.

Having obtained the 2D pose heatmaps, a directed GCN network is used to refine the potentially incomplete poses caused by occlusions or partially out-of-bounding box body parts, and two TCNs are used to estimate both person-centric 3D pose and camera-centric root depth based on a given sequence of 2D poses similar to [6]. As the TCN requires the input sequence of the same instance, a pose tracker [45] is used to track each instance in the input video. We also apply data augmentation in training our TCN so that it can handle occlusions [8].

3.2. Bottom-Up Network

Top-down methods perform estimation inside the bounding boxes, and thus are lack of global awareness of other persons, leading to difficulties to estimate poses in the camera-centric coordinates. To address this problem, we further propose a bottom-up network that processes multiple persons simultaneously. Since the bottom-up pose estimation suffers from human scale variations, we concatenate the heatmaps from our top-down network with the original input frame as the input of our bottom-up network. With the guidance of the top-down heatmaps, which are the results of the object detector and pose estimation based on the normalized boxes, the estimation of the bottom-up network will be more robust to scale variations. Our bottom-up network outputs four heatmaps : a 2D pose heatmap, ID-tag map, relative depth map, and root depth map. The 2D pose heatmap and ID-tag map are defined in the same way as in the previous section (3.1). The relative depth map refers to the depth map of each joint with respect to its root (pelvis) joint. The root depth map represents the depth map of the root joint.

In particular, the loss functions L_{hmap}^{BU} and L_{id}^{BU} for the heatmap and ID-tag map are similar to [33]. In addition, we apply the depth loss to the estimations of both the relative depth map h^{rel} and the root depth h^{root}. Please see supplementary material for example of the four estimated heatmaps from the bottom-up network. For N persons and K joints, the loss can be formulated as:

$$L_{depth} = \frac{1}{NK} \sum_n \sum_k |h_k(x_{nk}, y_{nk}) - d_{nk}|^2, \tag{2}$$

where h is the depth map and d is the ground-truth depth value. Note that, for pelvis (i.e., the root joint), the depth is a camera-centric depth. For other joints, the depth is relative with respect to the corresponding root joint.

We group the heatmaps into instances (i.e., persons), and retrieve the joint locations using the same procedure as in the top-down network. Moreover, the values of the camera-centric depth of the root joint z^{root} and the relative depth for the other joints z_k^{rel} are obtained by retrieving from the corresponding depth maps where the joints (i.e., root or others) are located. Specifically:

$$z_i^{root} = h^{root}(x_i^{root}, y_i^{root}) \tag{3}$$

$$z_{i,k}^{rel} = h_k^{rel}(x_{i,k}, y_{i,k}) \tag{4}$$

where i, k refer to the i_{th} instance and k_{th} joint, respectively.

3.3. Integration with Interaction-Aware Discriminator

Having obtained the results from the top-down and bottom-up networks, we first need to find the corresponding poses between the results from the two networks, i.e., the top-down pose P_i^{TD} and bottom-up pose P_j^{BU} belong to the same person. Note that P stands for camera-centric 3D pose throughout this paper.

Given two pose sets from bottom-up branch P^{BU} and top-down branch P^{TD}, we match the poses from both sets, in order to form pose pairs. The similarity of two poses is defined as:

$$\text{Sim}_{i,j} = \sum_{k=0}^{K} \min(c_{i,k}^{BU}, c_{j,k}^{TD}) \text{OKS}(P_{i,k}^{BU}, P_{j,k}^{TD}), \tag{5}$$

where:

$$\text{OKS}(x, y) = \exp(-\frac{d(x, y)^2}{2s^2\sigma^2}), \tag{6}$$

OKS stands for object keypoint similarity [49], which measures the joint similarity of a given joint pair. $d(x, y)$ is the Euclidean distance between two joints. s and σ are two controlling parameters. $\mathrm{Sim}_{i,j}$ measures the similarity between the i_{th} 3D pose P_i^{BU} from the bottom-up network and the j_{th} 3D pose P_j^{TD} from the top-down network over K joints. Note that both poses from top-down P^{TD} and bottom-up P^{BU} are camera-centric; thus, the similarity is measured based on the camera coordinate system. The $c_{i,k}^{BU}$ and $c_{j,k}^{TD}$ are the confidence values of joint k for 3D poses P_i^{BU} and P_j^{TD}, respectively. Having computed the similarity matrix between the two sets of poses P^{TD} and P^{BU} according to the $\mathrm{Sim}_{i,j}$ definition, the Hungarian algorithm [20] is used to obtain the matching results.

Once the matched pairs are obtained, we feed each pair of the 3D poses and the confidence score of each joint to our integration network. Our integration network consists of 3 fully connected layers, which outputs the final estimation.

Integration Network Training To train the integration network, we take some samples from the ground-truth 3D poses. We apply data augmentation: 1) random masking the joints with a binary mask M^{kpt} to simulate occlusions; 2) random shifting the joints to simulate the inaccurate pose detection; and 3) random zeroing one from a pose pair to simulate unpaired poses. The loss of the integration network is an L2 loss between the predicted 3D pose and its ground-truth:

$$L_{int} = \frac{1}{K} \sum_{k} |P_k - \tilde{P}_k|^2, \tag{7}$$

where K is the number of the estimated joints. P and $\tilde{P}$ are the estimated and ground-truth 3D poses, respectively.

Inter-Person Discriminator For training the integration network, we propose a novel inter-person discriminator. Unlike most existing discriminators for human pose estimation (e.g. [47, 7]), where they can only discriminate the plausible 3D poses of one person, we propose an interaction-aware discriminator to enforce the interaction of a pose pair is natural and reasonable, which not only includes the existing single-person discriminator, but also generalize to interacting persons. Specifically, our discriminator contains two sub-networks: D_1, which is dedicated for one person-centric 3D poses; and, D_2, which is dedicated for a pair of camera-centric 3D poses from two persons. We apply the following loss to train the network, which is formulated as:

$$L_{dis} = log(\tilde{C}) + log(1 - C) \tag{8}$$

where:

$$\begin{aligned} C &= 0.25(D_1(P^a) + D_1(P^b)) + 0.5D_2(P^a, P^b) \\ \tilde{C} &= 0.25(D_1(\tilde{P}^a) + D_1(\tilde{P}^b)) + 0.5D_2(\tilde{P}^a, \tilde{P}^b) \end{aligned} \tag{9}$$

where P^a, P^b are the estimated poses of person a and person b, respectively. $\tilde{P}$ are the estimated and ground-truth 3D poses, respectively.

3.4. Semi-Supervised Training

Semi-supervised learning is an effective technique to improve the network performance, particularly when the data with ground-truths are limited. A few works also explore to make use of the unlabeled data [4, 45, 51]. In our method, we apply a noisy student training strategy [50]. We first train a teacher network with the 3D ground-truth dataset only, and then use the teacher network to generate their pseudo-labels of unlabelled data, which are used to train a student network.

The pseudo-labels cannot be directly used because some of them are likely incorrect. Unlike in the noisy student training strategy [50], where data with ground-truth labels and pseudo-labels are mixed to train the student network by adding various types of noise (i. e., augmentations, dropout, etc), we propose two-consistency loss terms to assess the quality of the pseudo-labels, including the reprojection error and multi-perspective error [4, 36].

The reprojection error measures the deviation between the projection of generated 3D poses and the detected 2D poses. Since there are more abundant data variations in 2D pose dataset compared to 3D pose dataset (e.g., COCO is much larger compared to H36M), the 2D estimator is expected to be more reliable than its 3D counterpart. Therefore, minimizing a reprojection error is helpful to improve the accuracy of 3D pose estimation.

The multi-perspective error, E_{mp}, measures the consistency of the predicted 3D poses from different viewing angles. This error indicates the reliability of the predicted 3D poses. Based on the two terms, our semi-supervised loss, L_{SSL}, is formulated as,

$$L_{\mathrm{SSL}} = w(E_{rep} + E_{mp}) + L_{dis}, \tag{10}$$

where w is a weighting factor to balance the contribution of the reprojection and multi-perspective errors. In the training stage, w first focuses on easy samples and gradually includes the hard samples. The weight, w, is formulated as:

$$w = \mathrm{softmax}(\frac{E_{rep}}{r}) + \mathrm{softmax}(\frac{E_{mp}}{r}), \tag{11}$$

where r is the number of training epochs. More details regarding to the reprojection and multi-perspective errors and the self-training process are discussed in the supplementary material.

4. Experiment

Datasets We use MuPoTS-3D [29] and JTA [11] datasets to evaluate the camera-centric 3D multi-person pose estimation performance by following the existing methods [32, 10]

Method	AP_{25}^{root}	AUC_{rel}	PCK	PCK_{abs}
TD (w/o MP)	43.7	41.0	81.6	42.8
TD (w MP)	45.2	48.9	87.5	45.7
BU (w/o CH)	44.2	34.5	76.6	40.2
BU (w CH)	46.1	35.1	78.0	41.5
TD + BU (w/o MP,CH)	44.9	42.6	82.8	43.1
TD + BU (hard)	46.1	48.9	87.5	46.2
TD + BU (linear)	46.1	49.2	88.0	46.7
TD + BU (w/o PM)	46.0	48.6	85.5	45.3
TD + BU (IN)	**46.3**	**49.6**	**88.9**	**47.4**

Table 1. Ablation study on MuPoTS-3D dataset. TD, BU, MP, CH, IN, and PM stand for top-down, bottom-up, multi-person pose estimator, combined heatmap, integration network, and pose matching, respectively. Best in **bold**, second best underlined.

and their training protocols (i.e., train, test split). In addition, we use 3DPW [46] to evaluate person-centric 3D multi-person pose estimation performance following [17, 42]. We also perform evaluation on the widely used Human3.6M dataset [16] for person-centric 3D human pose estimation following [36, 47]. Details of the datasets information are in the supplementary material.

Implementation Details We use HRNet-w32 [41] as the backbone network for both multi-person pose estimator in the top-down and bottom-up networks. The top-down network is trained for 100 epochs on the COCO dataset [25] with the Adam optimizer and learning rate 0.001. The bottom-up network is trained for 50 epochs with the Adam optimizer and learning rate 0.001 on a combined dataset of MuCO [30] and COCO [25]. More details are in the supplementary material.

Evaluation Metrics Since the majority of 3D human pose estimation methods produce person-centric 3D poses, to be able to compare, we perform person-centric 3D human pose estimation. We use Mean Per Joint Position Error (MPJPE), Procrustes analysis MPJPE (PA-MPJPE), Percentage of Correct 3D Keypoints (PCK), and area under PCK curve from various thresholds (AUC_{rel}) following the literature [32, 36, 7]. Since we focus on 3D multi-person camera-centric pose estimation, we also use the metrics designed for evaluating performance in the camera coordinate system, including average precision of 3D human root location (AP_{25}^{root}) and PCK_{abs}, which is PCK without root alignment to evaluate the absolute camera-centric coordinates from [32], and F1 value following [10].

Ablation Studies Ablation studies are performed to validate the effectiveness of each sub-module of our framework. We validate our top-down network by using an existing top-down pose estimator (i.e., detection of one full-body joints) as a baseline, abbreviated as TD (w/o MP) to compare to our top-down network denoted as TD (w MP). We also validate our bottom-up network by using existing bottom-up heatmap estimation (i.e., estimate all person at the same scale) as a baseline, named BU (w/o CH) to compare to our bottom-up network, called BU (w CH). To evaluate our integration network, we use three baselines. The first is a straightforward integration by combining existing TD and BU networks. The second is hard integration, abbreviated TD + BU (hard), where the top-down person-centric pose is always used, plus the root depth from the bottom-up network. The third is linear integration, abbreviated TD + BU (linear), where the person-centric 3D pose from top-down is combined with its corresponding bottom-up one based on the confidence values of the estimated heatmap.

Method	AP_{25}^{root}	AUC_{rel}	PCK	PCK_{abs}
Rep	**46.3**	43.4	77.2	40.7
MP	**46.3**	32.2	72.8	29.5
Rep+dis	**46.3**	49.9	89.1	46.8
Rep+MP+dis	**46.3**	**50.6**	**89.6**	**48.0**

Table 2. Ablation study on MuPoTS-3D dataset. Rep, MP, and dis stand for reprojection, multi-perspective, and discriminator. Best in **bold**, second best underlined.

As shown in Table 1, we observe that our top-down network, bottom-up network, and integration network clearly outperform their corresponding baselines. Our top-down network tends to have better person-centric 3D pose estimations compared with our bottom-up network, because the top-down network benefits from not only multi-person pose estimator, also GCN and TCN that help to deal with inter-occluded poses. On the contrary, our bottom-up network achieves better performance for the root joint estimation, because it estimates the root depth based on a full image; while the root depth of top-down network is estimated based on an individual skeleton. Finally, our integration network demonstrates superior performance compared to hard or linear combining the poses from the top-down and bottom-up networks, which validates its effectiveness.

Other than validating our top-down and bottom-up networks, we also perform ablation analysis on our semi-supervised learning. We show the result of using reprojection loss, multi-perspective loss, reprojection loss with our discriminator, and reprojection & multi-perspective loss with discriminator in Table 2. We can see that the reprojection loss is more useful than the multi-perspective loss because it leverages the information from the 2D pose estimator, which is trained with 2D datasets with a large number of poses and environment variations. More importantly, we observe that our proposed interaction-aware discriminator makes the largest performance improvement compared with the other modules, demonstrating the importance of enforcing the validity of the interaction between persons.

Quantitative Evaluation To evaluate the performance for 3D multi-person camera-centric pose estimation in both indoor and outdoor scenarios, we perform evaluations on MuPoTS-3D as summarized in Table 3. The results show

that our camera-centric multi-person 3D pose estimation outperforms the SOTA [22] on PCK_{abs} by 2.3%. We also perform person-centric 3D pose estimation evaluation using PCK where we outperform the SOTA method [24] by 2.1%. The evaluation on MuPoTS-3D shows that our method outperforms the state-of-the-art methods in both camera-centric and person-centric 3D multi-person pose estimation as our framework overcomes the weaknesses of both bottom-up and top-down branches and at the same time benefits from their strengths.

Following recent work [10], we also perform evaluations on JTA, which is a synthetic dataset acquired from computer game, to further validate the effectiveness of our method for camera-centric 3D multi-person pose estimation. As shown in Table 4, our method is superior over the SOTA method [10] (e.g., our result shows 12.6% improvement on F1 value, $t = 0.4m$) on this challenging dataset where both inter-person occlusion and large person scale variation present, which again illustrate that our proposed method can handle these challenges in 3D multi-person pose estimation.

Human3.6M is widely used for evaluating 3D single-person pose estimation. As our method is focused on dealing with inter-person occlusion and scale variation, we do not expect our method performs significantly better than the SOTA methods. Table 5 summarizes the quantitative evaluation on Human3.6M where our method is comparable with the SOTA methods [19, 22] on person-centric 3D human pose evaluation metrics (i.e., MPJPE and PA-MPJPE).

3DPW is an outdoor multi-person 3D human shape reconstruction dataset. It is unfair to compare the errors between skeleton-based method with ground-truth defined on SMPL model [26] due to the different definitions of joints [44]. We run human detection on all frames and create an occlusion subset where the frames with the large overlay between persons are selected. The performance drop between the full testing test of 3DPW and the occlusion subset can effectively tell if a method can handle inter-person occlusion, which is shown in Table 6. We observe that our method shows the least performance drop from the testing set to the subset, which demonstrates our method is indeed more robust to inter-person occlusion.

Qualitative Evaluation Fig. 4 shows the comparison among a SOTA bottom-up method SMAP [53], our bottom-up branch, top-down branch, and full model. We observe that SMAP suffers from person scale variation where the person who is far from the camera is missing in frame 280 as well as inter-occlusion (e.g., frame 365 and 340). Our bottom-up branch is robust to scale variance, but fragile to the out-of-image poses as our discriminator is not used here (e.g., frame 365 and 330). Moreover, our top-down branch produces reasonable relative poses with the aid of GCN and TCNs. However, there exists error of camera-centric root depth in our top-down branch, because our top-down branch estimates root depth based on individual 2D poses and lacks global awareness (e.g., frame 280). Finally, our full model benefits from both branches and produces the best 3D pose estimations among these baselines.

We also provide results of the estimated 3D poses in

Group	Method	PCK	PCK_{abs}
Person-centric	Mehta et al. [29]	65.0	n/a
	Rogez et al., [39]	70.6	n/a
	Cheng et al. [8]	74.6	n/a
	Cheng et al. [7]	80.5	n/a
Camera-centric	Moon et al. [32]	82.5	31.8
	Lin et al. [24]	83.7	35.2
	Zhen et al. [53]	80.5	38.7
	Li et al. [22]	82.0	43.8
	Cheng et al. [6]	<u>87.5</u>	<u>45.7</u>
	Our method	**89.6**	**48.0**

Table 3. Quantitative evaluation on multi-person 3D dataset, MuPoTS-3D. Best in **bold**, second best <u>underlined</u>.

Method	$t = 0.4m$	$t = 0.8m$	$t = 1.2m$
[37] + [27] + [39]	39.14	47.38	49.03
LoCO [10]	<u>50.82</u>	<u>64.76</u>	<u>70.44</u>
Ours	**57.22**	**68.51**	**72.86**

Table 4. Quantitative results on JTA dataset. F1 values are reported based on different threshold t when the point is considered "true positive" when the distance from corresponding distance is less than t. Best in **bold**, second best <u>underlined</u>.

Group	Method	MPJPE	PA-MPJPE
Person-centric	Hossain et al., [13]	51.9	42.0
	Wandt et al., [47]*	50.9	38.2
	Pavllo et al., [36]	46.8	36.5
	Cheng et al., [8]	42.9	32.8
	Kocabas et al., [18]	65.6	41.4
	Kolotouros et al. [19]	n/a	<u>41.1</u>
Camera-centric	Moon et al., [32]	54.4	35.2
	Zhen et al., [53]	54.1	n/a
	Li et al., [22]	48.6	<u>30.5</u>
	Ours	**40.7**	**30.4**

Table 5. Quantitative evaluation on Human3.6M for normalized and camera-centric 3D human pose estimation. * denotes ground-truth 2D labels are used. Best in **bold**, second best <u>underlined</u>.

Dataset	Method	PA-MPJPE	δ
Original	Doersch et al. [9]	74.7	n/a
	Kanazawa et al. [17]	72.6	n/a
	Arnab et al. [1]	72.2	n/a
	Cheng et al. [7]	71.8	n/a
	Sun et al. [42]	69.5	n/a
	Kolotouros et al. [19]*	<u>59.2</u>	n/a
	Kocabas et al., [18]*	**51.9**	n/a
	Our method	62.9	n/a
Subset	Cheng et al. [7]	92.3	+20.5
	Sun et al. [42]	84.4	<u>+14.9</u>
	Kolotouros et al. [19]*	79.1	+19.9
	Kocabas et al., [18]*	**72.2**	+20.3
	Our method	<u>75.6</u>	**+12.7**

Table 6. Quantitative evaluation using PA-MPJPE on original 3DPW test set and its occlusion subset. * denotes extra 3D datasets were used in training. Best in **bold**, second best <u>underlined</u>.

Figure 4. Examples of results from our whole framework compared with different baseline results. First row shows the images from two video clips; second row shows the results from SMAP [53]; third row shows the result of of our bottom-up (BU) branch; fourth row shows the results of our top-down (TD) branch; last row shows the results of our full model. Wrong estimations are labeled with red circles.

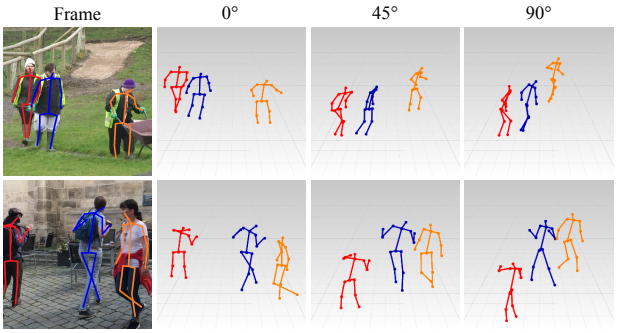

Figure 5. Qualitative results of the estimated 2D poses overlaying on input images and the estimated 3D poses visualized in novel viewpoints (virtual camera rotated by 0, 45, 90 degrees clockwise). Different colors are used for different persons in both 2D and 3D human poses for better visualization purpose.

novel viewpoints and the estimated 2D poses overlaid on input images as in Fig. 5 where our estimated camera-centric 3D poses visualized from different angles further validate the effectiveness of our method. Two failure cases are shown in Fig. 6 where the samples are taken from MPII dataset. The common failure cases are constant heavy occlusion (left) and unusual poses (right).

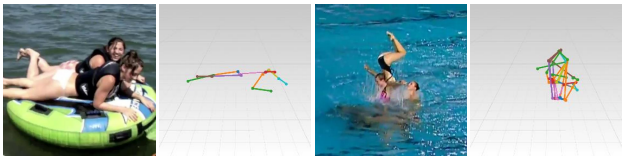

Figure 6. Two representative failure cases of our method.

5. Conclusion

We have proposed a novel method for monocular-video 3D multi-person pose estimation, which addresses the problems of inter-person occlusion and close interactions. We introduced the integration of top-down and bottom-up approaches to exploit their strengths. Our quantitative and qualitative evaluations show the effectiveness of our method compared to the state-of-the-art baselines.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

[1] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3395–3404, 2019. 7

[2] Lorenzo Bertoni, Sven Kreiss, and Alexandre Alahi. Monoloco: Monocular 3d pedestrian localization and uncertainty estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 1

[3] Yujun Cai, Liuhao Ge, Jun Liu, Jianfei Cai, Tat-Jen Cham, Junsong Yuan, and Nadia Magnenat Thalmann. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2272–2281, 2019. 3

[4] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019. 5

[5] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5386–5395, 2020. 2, 4

[6] Yu Cheng, Bo Wang, Bo Yang, and Robby T Tan. Graph and temporal convolutional networks for 3d multi-person pose estimation in monocular videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 4, 7

[7] Yu Cheng, Bo Yang, Bo Wang, and Robby T Tan. 3d human pose estimation using spatio-temporal networks with explicit occlusion training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8118–8125, 2020. 1, 2, 5, 6, 7

[8] Yu Cheng, Bo Yang, Bo Wang, Wending Yan, and Robby T Tan. Occlusion-aware networks for 3d human pose estimation in video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 723–732, 2019. 2, 4, 7

[9] Carl Doersch and Andrew Zisserman. Sim2real transfer learning for 3d human pose estimation: motion to the rescue. In *Advances in Neural Information Processing Systems*, pages 12929–12941, 2019. 2, 7

[10] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Compressed volumetric heatmaps for multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7204–7213, 2020. 2, 5, 6, 7

[11] Matteo Fabbri, Fabio Lanzi, Simone Calderara, Andrea Palazzi, Roberto Vezzani, and Rita Cucchiara. Learning to detect and track visible and occluded body joints in a virtual world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 430–446, 2018. 5

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 4

[13] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *ECCV*, pages 69–86. Springer, 2018. 2, 7

[14] Peiyun Hu and Deva Ramanan. Bottom-up and top-down reasoning with hierarchical rectified gaussians. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5600–5609, 2016. 3

[15] Gang Hua, Ming-Hsuan Yang, and Ying Wu. Learning to estimate human pose with data driven belief propagation. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 747–754. IEEE, 2005. 3

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2, 6

[17] Angjoo Kanazawa, Jason Y. Zhang, Panna Felsen, and Jitendra Malik. Learning 3d human dynamics from video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6, 7

[18] Muhammed Kocabas, Nikos Athanasiou, and Michael J Black. Vibe: Video inference for human body pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5253–5263, 2020. 7

[19] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2252–2261, 2019. 7

[20] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 5

[21] Paul Kuo, Dimitrios Makris, and Jean-Christophe Nebel. Integration of bottom-up/top-down approaches for 2d pose estimation using probabilistic gaussian modelling. *Computer Vision and Image Understanding*, 115(2):242–255, 2011. 3

[22] Jiefeng Li, Can Wang, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 7

[23] Miaopeng Li, Zimeng Zhou, and Xinguo Liu. Multi-person pose estimation using bounding box constraint and lstm. *IEEE Transactions on Multimedia*, 21(10):2653–2663, 2019. 3

[24] Jiahao Lin and Gim Hee Lee. Hdnet: Human depth estimation for multi-person camera-space localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 7

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 6

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):248, 2015. 7

[27] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2, 7

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. Xnect: Real-time multi-person 3d motion capture with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 39(4):82–1, 2020. 1, 2, 3

[29] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *2018 International Conference on 3D Vision (3DV)*, pages 120–130. IEEE, 2018. 1, 5, 7

[30] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d pose estimation from monocular rgb. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE, sep 2018. 6

[31] Dushyant Mehta, Srinath Sridhar, Oleksandr Sotnychenko, Helge Rhodin, Mohammad Shafiei, Hans-Peter Seidel, Weipeng Xu, Dan Casas, and Christian Theobalt. Vnect: Real-time 3d human pose estimation with a single rgb camera. *ACM Transactions on Graphics (TOG)*, 36(4):44, 2017. 2

[32] Gyeongsik Moon, Juyong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *The IEEE Conference on International Conference on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6, 7

[33] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in neural information processing systems*, pages 2277–2287, 2017. 4

[34] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3467–3475. IEEE, 2017. 2

[35] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7307–7316, 2018. 2

[36] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7753–7762, 2019. 2, 5, 6, 7

[37] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 7

[38] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3433–3441, 2017. 2

[39] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2, 7

[40] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010. 2

[41] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 6

[42] Yu Sun, Yun Ye, Wu Liu, Wenpeng Gao, YiLi Fu, and Tao Mei. Human mesh recovery from monocular images via a skeleton-disentangled representation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5349–5358, 2019. 6, 7

[43] Wei Tang, Pei Yu, and Ying Wu. Deeply learned compositional models for human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 190–206, 2018. 3

[44] Shashank Tripathi, Siddhant Ranade, Ambrish Tyagi, and Amit Agrawal. Posenet3d: Unsupervised 3d human shape and pose estimation. *arXiv preprint arXiv:2003.03473*, 2020. 7

[45] Rafi Umer, Andreas Doering, Bastian Leibe, and Juergen Gall. Self-supervised keypoint correspondences for multi-person pose estimation and tracking in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 4, 5

[46] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, pages 614–631, 2018. 6

[47] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5, 6, 7

[48] Sheng Wang, Haizhou Ai, Takayoshi Yamashita, and Shihong Lao. Combined top-down/bottom-up human articulated pose estimation using adaboost learning. In *2010 20th International Conference on Pattern Recognition*, pages 3670–3673. IEEE, 2010. 3

[49] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 466–481, 2018. 5

[50] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10687–10698, 2020. 5

[51] Xiangyu Xu, Hao Chen, Francesc Moreno-Noguer, László A Jeni, and Fernando De la Torre. 3d human shape and pose from a single low-resolution image with self-supervised learning. *arXiv preprint arXiv:2007.13666*, 2020. 5

[52] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5255–5264, 2018. 2

[53] Jianan Zhen, Qi Fang, Jiaming Sun, Wentao Liu, Wei Jiang, Hujun Bao, and Xiaowei Zhou. Smap: Single-shot multi-person absolute 3d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 7, 8