PI-Net: Pose Interacting Network for Multi-Person Monocular 3D Pose Estimation

Wen Guo[1], Enric Corona[2], Francesc Moreno-Noguer[2], Xavier Alameda-Pineda[1]
[1]Inria, Univ. Grenoble Alpes, CNRS, Grenoble INP, LJK, 38000 Grenoble, France
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[1]{wen.guo,xavier.alameda-pineda}@inria.fr, [2]{ecorona,fmoreno}@iri.upc.edu

Abstract

Recent literature addressed the monocular 3D pose estimation task very satisfactorily. In these studies, different persons are usually treated as independent pose instances to estimate. However, in many every-day situations, people are interacting, and the pose of an individual depends on the pose of his/her interactees. In this paper, we investigate how to exploit this dependency to enhance current – and possibly future – deep networks for 3D monocular pose estimation. Our pose interacting network, or PI-Net, inputs the initial pose estimates of a variable number of interactees into a recurrent architecture used to refine the pose of the person-of-interest. Evaluating such a method is challenging due to the limited availability of public annotated multi-person 3D human pose datasets. We demonstrate the effectiveness of our method in the MuPoTS dataset, setting the new state-of-the-art on it. Qualitative results on other multi-person datasets (for which 3D pose ground-truth is not available) showcase the proposed PI-Net. PI-Net is implemented in PyTorch and the code will be made available upon acceptance of the paper.

1. Introduction

Monocular 3D multi-person human pose estimation aims at estimating the 3D joints of several people from a single RGB image. This problem attracts great research and industrial interests, as it would make possible a number of applications in many different fields including the entertainment industry, sports technology, physical therapy and medical diagnosis. Recent works on multi-person human pose estimation usually regard different people as independent instances and estimate the poses one by one in separate bounding boxes in top-down methods. This makes all these approaches agnostic about the context information and specifically about the presence of other people [4, 13, 26, 33, 34, 36, 44, 45, 50]. However, when people interact, the pose and motion of every person is typically dependent and correlated to the body posture of the people he/she is interacting with.

While context information has been shown to be useful in tasks such as object detection [2, 14, 39], motion prediction [11] or affordance estimation [12], to the best of our knowledge, it has not been well developed before in a body pose estimation. In this paper, we investigate how these dependencies can be used to boost the performance of off-the-shelf architectures for 3D human pose estimation.

Concretely, we propose a pose interacting network, PI-Net, which is fed with the 3D pose of a person of interest and an arbitrary number of body poses from other people in the scene, all of them computed with a context agnostic pose detector. These poses are potentially noisy, both in their absolute position in space as in the specific representation of the body posture. PI-Net is built using a recurrent network with a self-attention module that encodes the contextual information. Since it is unclear how to rank the contextual information, that is the pose of other persons, regarding the potential impact on the pose refinement pipeline, we make the very straightforward assumption that the potential of a person to refine the pose of the person-of-interest, is inversely proportional to the square of the distance between them.

We thoroughly evaluate our approach on the MuPoTS dataset [34], and using the initial detections of 3DMPPE [36], the current best performing approach on this dataset. PI-Net exhibits consistent improvement of the pose estimates provided by 3DMPPE in all sequences of the dataset, becoming thus, the new state-of-the-art (see one example in Fig. 1). Interestingly, note that PI-Net can be used as drop-in replacement for any other architecture that estimates 3D human pose. Additionally, the size of the network we propose is relatively small ($3.41M$ training parameters, while the baseline model has $36.25M$ parameters), enabling efficient training and introducing a marginal computational cost at test. Testing on one Geforce1070, PI-net just cost 0.007s on refining one person while the

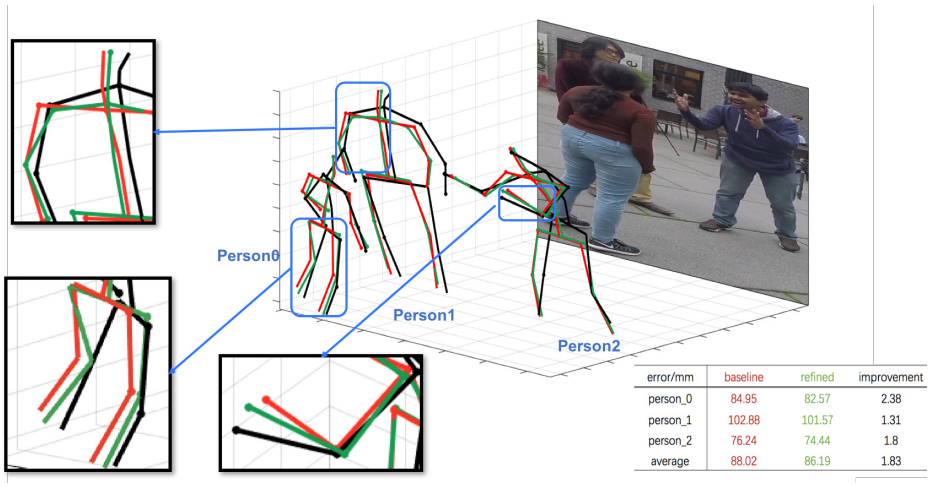

Figure 1: **PI-Net peformance.** An example of testing on MuPoTS dataset. Poses refined by PI-Net (in green) are closer to the ground truth (in black) than the baseline (in red). We zoom-in to several parts to clearly appreciate the difference. The error before and after PI-Net refinement for each person is shown in the table. The average 3D joint error for this example is reduced from 88.02 mm to 86.19 mm.

baseline cost 0.038s for detecting one root-centered pose and also extra time on obtaining the bounding boxes and roots. Our method is lightweight and consistently improves the baseline.

2. Related Work

2.1. 3D Single-person pose estimation

Deep learning methods for single-person 3D pose estimation follow two different strategies. On one hand, there are algorithms that directly learn the mapping from image features to 3D poses [10, 29, 32, 41, 43]. For instance, [29] propose a joint model for body part detectors and pose regression. Pavlakos *et al.* [41] introduce a U-Net architecture to recover joint-wise 3D heatmaps. Sun *et al.* [48] build a regression approach using a bone-based representation that enforces human pose structure. In [49], a differentiable soft-argmax operation is used for efficiently training a hourglass network.

Another line of work focuses on recovering 3D human pose from 2D image features by using models that enforce consistency between 3D predicted poses and 2D observations [5, 37, 51]. For instance, Bogo *et al.* [5] fit a human body parametric model by minimising the distance between the projection of the 3D estimation and the 2D predicted joints. Moreno-Noguer [37] propose to infer 3D pose via distance matrix regression. Yang *et al.* [54] use an adversarial approach to ensure that estimated poses are antropomorphic.

2.2. 2D multi-person pose estimation

There are two main approaches for multi-person pose estimation, top-down [8, 30, 47, 53] and bottom-up models [6, 7, 38, 42]. On the former, a human detector first estimates the bounding boxes containing the person. Each detected area is cropped and fed into the pose estimation network. The later also follows a two-stage pipeline, where a model first estimates all human body keypoints, and then groups them into each person using clustering techniques.

Cao *et al.* [6, 7] propose a real-time bottom-up method using Part Affinity Fields to group joints of different person. The efficiency of these bottom-up approaches makes them very appropriate to be used as a backbone for later lifting the 2D joints to 3D [13, 40]. The performance of bottom-up methods has been recently improved by to-down strategies. Xiao *et al.* [53] use ResNet [20] as encoder and several deconvolutional layers as decoder to formulate a simple but effective baseline. Sun*et al.* [47] connect the high-to-low resolution convolution streams in parallel to maintain richer semantic information. Chen *et al.* [8] use a cascade pyramid network to refine the hard keypoints of the initial estimated results.

2.3. 3D Multi-person pose estimation and contextual information

Similar to their 2D counterparts, 3D multi-person poses estimation methods can be split into top-down [26, 36, 44, 45, 52] and bottom-up [33, 34, 56] approaches. Mehta *et al.* [33, 34] follow a bottom-up strategy, by first estimating three occlusion-robust location-maps [35] and then mod-

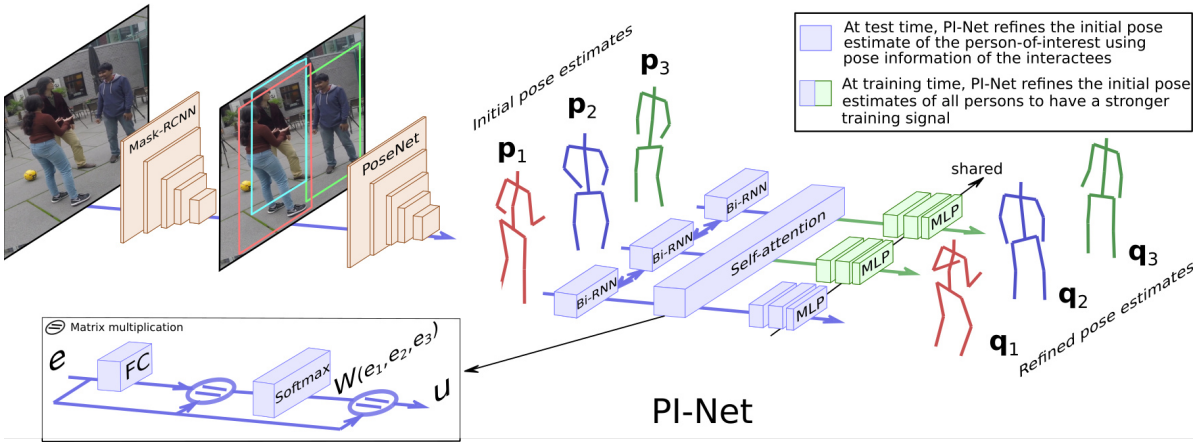

Figure 2: **PI-Net Architecture**. Mask-RCNN [19] and PoseNet [36] are used to extract the initial pose estimates $\mathbf{p}_1, \ldots, \mathbf{p}_N$. These estimates are fed into PI-Net, composed of three main blocks: Bi-RNN, Self-attention and the shared fully-connected layers. The output of PI-Net refines the initial pose estimates by exploiting the pose of the interactees, yielding $\mathbf{q}_1, \ldots, \mathbf{q}_N$.

eling the association between body keypoints using Part Affinity Fields [7]. Zanfir *et al*. [56] formalize the problem of localizing and grouping people as a binary linear integer program and solve it by integrating a limb scoring model.

Rogez *et al*. [44, 45], in contrast, propose a top-down approach, where first, each person 2D bounding box is classified into one of the anchor clustered 3D poses. These poses are then refined in a coarse-to-fine manner. Moon *et al*. [36] propose an architecture that simultaneously predicts the 3D absolute position of the root joint and reconstructs the relative 3D body pose of multiple people. However, despite the fact that these works estimate the body pose of an arbitrary number of people, each person is processed using an independent pipeline that does not take into account the interactions between the rest of people or other contextual information.

Recently, some works begin to pay attention to using contextual information in 3D pose estimation problem by integrating scene constraints [55] or considering the depth-order to resolve the overlapping problem [24, 28]. Jiang *et al*. [24] propose a depth ordering-aware loss to consider the occlusion relationship and interpenetration of people in multi-person scenarios. Li *et al*. [28] divide human relations into different levels and define 3 corresponding losses to tell if the orders of different people or different joints are correct or not. Though contextual information is considered in these works, they do not really explore the interaction relations between different people in the same activity. More recently, Fieraru *et al*. [16] proposed a new dataset of human interactions with several daily interaction scenarios and proposed a framework based on contact detection over model surface regions, but this dataset is not released yet.

In this paper, we propose a method that can be used in combination with the current state-of-the-art model [36] and boost its performance by looking at the whole group of humans. The proposed model is flexible and can be stacked after any 3D pose estimation model, independently of it being top-down or bottom-up.

3. PI-Net for Multi-Person Pose Estimation

Our goal is to exploit the interaction information between N people so as to improve the estimation of their pose. We assume the existence of an initial 3D pose estimate $\mathbf{p}_n \in \mathbb{R}^{J\times 3}$ of person $n = 1, \ldots, N$, where J is the number of estimated joints, *e.g*. obtained from 3DMPPE [36]. All the N poses are in absolute camera coordinates.

Formally, our goal is to improve the initial pose estimates, taking into account the pose of other people:

$$[\mathbf{q}_1, \ldots, \mathbf{q}_N] = \mathbf{\Pi}(\mathbf{p}_1, \ldots, \mathbf{p}_N), \tag{1}$$

where $\mathbf{q}_n \in \mathbb{R}^{J\times 3}$ denotes the pose of person n improved with the information of the poses of the interactees.

While the idea is very intuitive, the research question is how to design PI-Net (i.e. $\mathbf{\Pi}$) so that it satisfies the following desirable criteria. Firstly, it shall work in environments with different number of people N, and not fixed to a particular scenario. Secondly, the interaction information can be efficiently exploited and learned using publicly available datasets. Finally, it has to be generic enough to work with *any* 3D monocular multi-person pose estimator.

3.1. Pipeline of PI-Net

Naturally, the fact that the number of people N is unknown in advance, points us towards the use of recurrent neural networks. Such RNN should input the poses estimated by a generic pose estimator, and embed the pose information into a representation learned specifically to take the cross-interactions into account. Without loss of generality, let us assume that the person-of-interest is $n = 1$, and hence the pose to refine is $\mathbf{p}_1$. We consider using a bi-directionnal RNN, whose first input is $\mathbf{p}_1$, and then the rest of initial poses are provided in a given order (see below). Our intuition for using a Bi-RNN is the following. During the forward pass, and since the first input is $\mathbf{p}_1$, the network can use the information in $\mathbf{p}_1$ to extract the features of the other poses that will best refine $\mathbf{p}_1$. In the backward pass, the network accumulates all this information back to $\mathbf{p}_1$, obtaining:

$$\mathbf{e}_1 = \text{Bi-RNN}(\mathbf{p}_1, \dots, \mathbf{p}_N). \quad (2)$$

The learned embedding $\mathbf{e}_1 \in \mathbb{R}^{N \times E}$is supposed to contain the crucial information from all other poses to refine the pose of the person-of-interest (1 in our example), but not only. Indeed, given that a priori we do not know which persons would be more helpful in refining the pose of interest, the computed embedding $\mathbf{e}_1$ could contain information that is not exploitable to refine the pose. In order to take this phenomenon into account, we soften the requirements of the Bi-RNN through the use of an attention mechanism as shown in Figure 2 (bottom-left zoom). Such attention mechanism aims to improve each embedding by combining information from the embeddings of other persons. To do so, we compute a matrix of attention weights:

$$\mathbf{W} \in \mathbb{R}^{N \times N}, \qquad \mathbf{W}_{nm} = \mathbf{e}_n^\top (\mathbf{A}_{\text{ATT}} \mathbf{e}_m + \mathbf{b}_{\text{ATT}})), \quad (3)$$

that is then normalised with a row-wise soft-max operation. $\mathbf{A}_{\text{ATT}}$ and $\mathbf{b}_{\text{ATT}}$ are attention parameters to be learned. The self-attention weights $\mathbf{W}$ encoding the residual interaction not captured by the Bi-RNN are used to update the embedding vector $\mathbf{u}_1 = \mathbf{W}\mathbf{e}_1$. Finally, the updated embedding is feed-forwarded through a few fully connected layers, obtaining the final refined pose $\mathbf{q}_1$. While, at test time the self-attention and fully-connected layers are used only for the person-of-interest, at training time we found it is useful to apply these two operations to all poses, and back-propagate the loss associated to everyone. This strategy eases the training. The overall pipeline depicting of PI-Net is shown in Figure 2.

3.2. Interaction Order

In the previous section we assumed that the order in which the initial pose estimates $\mathbf{p}_n$ were presented to the Bi-RNN was given. Although there is no principled rule to define the ordering, there are some requirements. For a given person n, the sequence of poses presented to the network $\mathbf{p}_{\rho_n(1)}, \dots, \mathbf{p}_{\rho_n(N)}$ has two constraints: (i) each pose is presented only once and (ii) the first pose is the one to be refined, i.e. $\rho_n(1) = n$. Intuitively, the order should represent the relevance: the more useful $\mathbf{p}_m$ is to refine $\mathbf{p}_n$, the closer $\mathbf{p}_m$ should be to $\mathbf{p}_n$ in the input sequence, i.e. the smaller $\rho_n(m)$ should be. Because finding the optimal permutation is a complex combinatorial optimisation problem for which there is no ground-truth, we opt for assuming that the relevance is highly correlated to the physical proximity between interactees. Therefore, the closer person m is to person n, the smaller should $\rho_n(m)$ be. With this rule we order the initial pose estimates to be fed to the Bi-RNN.

We also consider of using Graph Convolutional Network [25] to model the interaction between different person. Considering a pair of input persons, the node of the graph represents the coordinate of all the joints of these two people, and the adjacency matrix learned from the input represents interaction between these joints. This strategy does not provide any performance increase, the results will be discussed in Section 4.4.

3.3. Network Architecture

In order to build and train our PI-Net, we first extract the initial poses using [36]. In the baseline, Mask-RCNN is used to detect the people present in the image. After that, the keypoint detector is applied to each image to detect the root-based poses and then project them into absolute camera coordinates. This keypoint detector is based on ResNet50 and 3 addition deconvolutional layers, following [49]. The set of keypoints for each person in camera coordinates $\mathbf{p}_n$, is therefore obtained. Note that this regressor gives all J person joints, despite of partial occlusions, the corresponding occluded joints are hallucinated.

These initial pose estimates are then normalised with their mean and standard deviation, thus obtaining the input pose estimates of our PI-Net, $\{\mathbf{p}_1, \dots, \mathbf{p}_N\}$. For each person n, we feed the PI-Net with the sequence of poses in the order appropriate for person n (see Section 3.2). The output $\mathbf{q}_n$ of PI-Net is the refined pose for person n. PI-Net is trained with the L_1 loss between the refined poses and the ground-truth in 3D camera coordinates, added for all detected persons in the training image.

The Bi-RNN is implemented using three layers of gated recurrent units (GRU [9]). The the self-attention layer provides a straightforward way to account for person pose interactions. After applying attention, the updated embedding goes through three fully connected layers to output the refined 3D pose in camera coordinates. These three fully connected layers are shared by all N poses. Consequently, the proposed PI-Net can be trained and evaluated using images with different number of people.

Table 1: Sequence-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, the fourth method and our model is tested with ground truth bounding boxes and roots. Higher value means better performance.

Sequence	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	AVG
Accuracy for all ground truths																					
LCR [44]	67.7	49.8	53.4	59.1	67.5	22.8	43.7	49.9	31.1	78.1	50.2	51.0	51.6	49.3	56.2	66.5	65.2	62.9	66.1	59.1	53.8
Singleshot [34]	81.0	60.9	64.4	63.0	69.1	30.3	65.0	59.6	64.1	83.9	68.0	68.6	62.3	59.2	70.1	80.0	79.6	67.3	66.6	67.2	66.0
Xnect [33]	88.4	65.1	68.2	72.5	76.2	46.2	65.8	64.1	75.1	82.4	74.1	72.4	64.4	58.8	73.7	80.4	84.3	67.2	74.3	67.8	70.4
LCR++ [45]	87.3	61.9	67.9	74.6	78.8	48.9	58.3	59.7	78.1	89.5	69.2	73.8	66.2	56.0	74.1	82.1	78.1	72.6	73.1	61.0	70.6
PandaNet [3]	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	72.0
3DMPPE [36]	93.2	75.6	80.3	81.5	84.6	75.3	**84.5**	69.3	90.1	92.0	81.0	81.0	73.4	73.5	81.8	89.6	88.4	84.3	74.5	70.6	81.2
PI-Net (ours)	**93.5**	**77.4**	**82.0**	**82.9**	**87.2**	**75.9**	84.0	**71.5**	**90.2**	**92.2**	**82.5**	**82.9**	**74.7**	**75.7**	**83.6**	**91.4**	**90.6**	**86.0**	**74.9**	**71.1**	**82.5**
Accuracy only for matched ground truths																					
LCR [44]	69.1	67.3	54.6	61.7	74.5	25.2	48.4	63.3	69.0	78.1	53.8	52.2	60.5	60.9	59.1	70.5	76.0	70.0	77.1	81.4	62.4
Singleshot [34]	81.0	65.3	64.6	63.9	75.0	30.3	65.1	61.1	64.1	83.9	72.4	69.9	71.0	72.9	71.3	83.6	79.6	73.5	78.9	90.9	70.8
LCR++ [45]	88.0	73.3	67.9	74.6	81.8	50.1	60.6	60.8	78.2	89.5	70.8	74.4	72.8	64.5	74.2	84.9	85.2	78.4	75.8	74.4	74.0
Xnect [33]	88.4	70.4	68.3	73.6	82.4	46.4	66.1	83.4	75.1	82.4	76.5	73.0	72.4	73.8	74.0	83.6	84.3	73.9	85.7	90.6	75.8
3DMPPE [36]	**93.9**	83.0	80.3	81.5	85.4	75.3	84.5	77.2	90.1	92.0	81.0	81.0	74.3	76.0	81.8	89.6	88.4	84.3	75.5	76.2	82.6
PI-Net (ours)	**93.9**	**85.0**	**81.5**	**83.0**	**88.9**	**75.6**	**84.7**	**78.0**	**90.4**	**92.2**	**82.5**	**82.6**	**76.0**	**77.6**	**83.5**	**91.5**	**90.5**	**85.9**	**75.7**	**78.5**	**83.9**

3.4. Implementation details

We use PoseNet of 3DMPPE [36] to generate our input 3D human pose. This model is trained on large-scale training data which includes H3.6M single-person 3D dataset [23], MPII [1] and COCO 2D dataset [31], MuCo multi-person 3D dataset [34], and extra synthetic data. PI-Net is trained on 33.4k composited MuCo data, which is contained in the training data of the baseline model. This ensures that the improvement of PI-Net comparing with the baseline model is not caused by adding extra training data.

In terms of dimensions, 3DMPPE [36] outputs $J = 17$ joints in 3D, the hidden recurrent layers are of dimension 256, and the Bi-RNN outputs an embedding vector of dimension $E = 512$. We train our PI-net using Adam optimization and the *poly learning rate policy* [57], with initial learning rate of 1e-5, final learning rate of 1e-8, and power of 0.9, for 25 epochs. Batch size is set to 4.

When testing on an image with n instances, we test for n independent times, each time with a different ordering, and just retain the first person in each case.

4. Experiments

We next describe the experiment section, which includes a description of the datasets, baselines and evaluation metrics. We then provide a quantitative and qualitative evaluation and comparison to state-of-the-art approaches. We finalize this section with an exhaustive ablation study of the PI-Net architecture and hyperparameters.

4.1. Datasets

MuCo-3DHP dataset and MuPoTS-3D dataset. Most experiments discussed below are performed using these two well-known datasets. They were initially introduced by Mehta et al. [34] and are typically used as train set and test set respectively, for the task of multi-person 3D human pose estimation. MuCo-3DHP is a multi-person 3D human pose dataset. Our PI-Net is trained on $33.4k$ MuCo images with $80.7k$ instances, without any other extra data. MuPoTS-3D test set includes 8320 images with $23k$ instances in 20 real scenes (5 indoor scenes and 15 outdoor scenes). Each scene contains from 200 to 800 frames extracted from a video, with 2 or 3 people performing a certain common activity such as talking, shaking hands or doing sports. These two datasets are annotated using COCO format and provide both 2D image coordinates and 3D camera coordinates for each body joint.

COCO dataset. We also perform qualitative results using the COCO dataset. This is a large-scale multi-person human pose dataset and, even though it just provides 2D ground truth labels, it depicts challenging scenes with a large number of people performing very diverse actions. In particular, we use examples from the COCO val2017 subset [19].

4.2. Baseline and Evaluation metrics

Our pipeline is capable of refining the poses estimated by any multi-person pose algorithm, independently of the strategy it uses. Given these initially estimated poses we refine them leveraging on the contextual information. In this paper, we use the recent 3DMPPE [36] as a baseline and demonstrate both quantitative and qualitative improve-

Table 2: PA MPJPE (top) and MPJPE (bottom) comparisons of PI-net with the state-of-the-art method [36] used as our baseline on the MuPoTS dataset. The average value indicated image-wise average. Ground truth bounding boxes and roots are used for testing. Lower value means better performance.

Sequence	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15	S16	S17	S18	S19	S20	AVG
PA MPJPE (mm)																					
3DMPPE [36]	67.7	102.6	82.7	82.5	79.8	**91.1**	70.8	110.1	72.8	63.5	88.6	79.6	105.1	110.5	77.5	72.2	73.3	86.8	91.9	120.0	88.4
PI-Net (ours)	**65.8**	**97.7**	**82.2**	**82.4**	**77.7**	91.6	**68.6**	**106.3**	**70.0**	**60.5**	**88.0**	**77.7**	**102.3**	**106.6**	**75.5**	**70.2**	**71.5**	**83.7**	**88.9**	**112.6**	**85.79**
MPJPE (mm)																					
3DMPPE [36]	90.9	159.3	121.8	113.5	107.8	**121.1**	113.8	138.2	99.7	98.4	119.6	115.4	143.7	151.7	111.7	101.8	105.6	115.8	140.7	187.7	126.0
PI-Net (ours)	**87.3**	**151.3**	**117.1**	**109.9**	**103.9**	**121.1**	**108.7**	**133.9**	**95.8**	**93.0**	**117.0**	**112.2**	**141.1**	**146.2**	**108.0**	**98.0**	**102.5**	**111.8**	**136.2**	**178.4**	**121.7**

Table 3: Joint-wise 3DPCK comparison with state-of-the-art methods on the MuPoTS-3D dataset. The first three methods show the reported results in the corresponding paper, the fourth method and our model is tested with ground truth bounding boxes and roots. All ground truths are used for evaluation. Higher value means better performance.

Method	Hd.	Nck.	Sho.	Elb.	Wri.	Hip	Kn.	Ank.	Avg
LCR [44]	49.4	67.4	57.1	51.4	41.3	84.6	56.3	36.3	53.8
single-shot [34]	62.1	81.2	77.9	57.7	47.2	**97.3**	66.3	47.6	66.0
3DMPPE [36]	**78.4**	**91.9**	83.1	79.7	67.0	93.9	84.3	**75.3**	81.2
PI-Net (ours)	78.3	91.8	**87.8**	**81.9**	**68.5**	94.2	**85.3**	74.8	**82.5**

ments. Note that previous state-of-art works such as PandaNet [3] or SingleShot [34] do not provide codes either for training or testing, and hence, we could not use them as backbones. The baseline [36] consists of 3 main steps. Firstly, 2D bounding boxes of humans are detected using Mask-RCNN [19]. For each detection, a deep network refines the coarse root 3D coordinates obtained from camera calibration parameters and, finally, a fully convolutional network [49] predicts root-relative 3D pose. Using the 3D root position, all poses can be represented in a common camera-coordinates reference.

We evaluate the performance of all methods by reporting the percentage of keypoints detected by the network that are within 150mm or less from the ground truth labels (3DPCK@150mm). This is the usual evaluation metric on the MuPOTS-3D test set [3, 33, 34, 36, 44, 45].

Notice that the 3DPCK metric depends greatly on the chosen threshold, for completeness, we also provide MPJPE and PA-MPJPE metrics to evaluate the performances. MPJPE indicates mean-per-joint-position error after root alignment with the ground truth [23], and PA-MPJPE denotes MPJPE after Procrustes Alignment [18]. Lower MPJPE and PA-MPJPE indicates better performance.

4.3. Main results

Quantitative results on MuPoTS-3D testset. We report results of PI-Net on the MuPoTS-3D dataset in Table 1, and compare to current state-of-the-art methods. Our results are obtained using the model depicted in Fig. 2, which uses a bidirectional 3-GRU recurrent layer, followed by a self-attention layer. We provide results after root alignment with the ground-truth poses, on the two strategies usually used on the MuPoTs datasets. In table 1, the top-rows *Accuracy for all ground truths* evaluates all annotated persons, and the bottom rows *Accuracy only for matched ground truths* evaluates only predictions matched to annotations by their 2D projections with the 2D ground truths. We got improvements on both of the two strategies. PI-Net outperforms all previous models and improves the state-of-the-art by 1.3% 3DPCK@150mm on average. The improvement is consistent and shows a boost in performance for the majority of actions, setting a new state-of-the-art on the MuPoTS-3D dataset. Interestingly, we observe that the largest improvements are produced in those actions that require harmony and certain synchronization between people, such as practicing Taekwondo (S2) or playing a ball together (S14). We use ground truth bounding box and roots to test the baseline, so the root-relative result is comparable with the absolute result here. To avoid the redundancy, we only report root-relative results, which is widely reported in the previous works, for the comparison with the state-of-the-art methods.

Table 2 shows the comparison of sequence-wise performance using MPJPE with root alignment and PA-MPJPE with further rigid alignment. Testing our model on the MuPoTS test dataset, we reduced the MPJPE error and PA-MPJPE error by 2.6mm and 4.3mm on average, respectively, in comparison with the baseline results [36]. Again, results are consistent across different tasks.

Table 3 shows a joint-wise comparison with state-of-art

Table 4: Comparison of different input orders. *Intuitive* is the one described in Section 3.2, from near to far. *Inverse* is the opposite. *Random* means in random order.

Order	PA MPJPE (mm)	MPJPE (mm)
Reverse	86.09	122.23
Random	85.87	121.88
Intuitive	**85.79**	**121.7**

methods using 3DPCK@150mm after root alignment with ground truths. While we achieve similar performance with [36] in head, neck and hip, our method consistently outperforms the rest of joints on arms and legs (shoulder, elbow, wrists and knees). Arguably, the joints on the torso have little influence on the interaction between people, which comes mostly through the limbs, for example hands and legs. Hence, it is reasonable that using the context information to refine 3D pose predictions gives the most significant boost in these joints.

Finally, it is worth pointing out that the results for all previous approaches reported in Tables 1, 2 and 3 are those of the respective papers. For 3DMPPE [36], however, we tested on ground-truth bounding boxes and roots to reported these results.

Qualitative results on COCO. Figure 3 shows qualitative results on COCO dataset, for which 3D ground truths are not available. We also include (bottom-right) a failure case, caused by a misdetection of the baseline. This is maybe the major limitation of PI-Net, which is designed to refine poses, but so far, we have not integrated any module to deal with large deviations on the input poses.

4.4. Ablation Study

We next provide further analysis of the PI-Net architectural design and discuss/interpret the predicted adjacency matrix obtained in the self-attention layers.

Effect of the Input Order. Table 4 shows the effect of using different strategies to establish the ordering of the detected people fed into the Bi-RNN layer. We consider three different order: (i) a random ordering, (ii) our approach where we select the person of interest followed by people in order of proximity, and (iii) the inverse approach that person further away is firstly fed into the network. To estimate the distance between people, we compute the distances between the root coordinates of the input people to the target person.

Even though the number of people in images of MuPoTS dataset is relatively small and therefore the results would not differ greatly, the ordering in which every person's information is processed has an effect in the performance of the model. As shown in the Table, the ordering we use provides the best performance and the inverse one results the worst. This demonstrates the importance of taking context into account.

Table 5: Importance of self-attention and bidirectionality (RNN). PI-Net uses a bidirectional RNN followed by a self-attention layer. We evaluate the impact of each of these choices: w/o Att. when removing attention, w/o Bi. considering standard RNN.

Method	PA MPJPE(mm)	MPJPE(mm)
PI-Net w/o Att., w/o Bi.	86.69	122.7
PI-Net w/o Bi.	86.42	123.10
PI-Net w/o Att.	85.92	122.01
PI-Net	**85.79**	**121.7**

Table 6: Ablating the unit of the interaction network: None [36], Graph Convolutional Networks (GCN); LSTM and Gated Recursive Units (GRU), with (2,3,4) layers.

Interaction	PA MPJPE (mm)	MPJPE (mm)	# Par.
None [36]	88.36	126.0	133M
GCN	88.67	126.3	34M
2 LSTM	86.45	122.5	2.78M
3 LSTM	86.17	122.3	4.36M
4 LSTM	86.32	121.7	5.93M
2 GRU	86.27	122.2	2.23M
3 GRU	**85.79**	**121.7**	3.41M
4 GRU	85.96	122.2	4.59M

Effect of self-attention and bidirectional RNN. In Table 5 we analyse the effect of using the self-attention layer, which confirms that it helps to boost the performance. We also study the attention weights predicted by the self-attention layer. These weight are, as expected, large at the diagonal, which corresponds to the self-interaction. The larger the distance between two person is, the smaller the weights tend to be. Table 5 also compares our approach which employs Bi-RNN with a standard (not bidirectional) RNN. The ablation of the recurrent unit is done later one. Bi-RNN reduces 0.69mm the MPJPE error and 0.77mm the PA-MPJPE error, while the self-attention layers gives an extra improvement of 0.31mm on MPJPE and 0.13mm on PA-MPJPE.

Interaction unit. In Table 6 we report results using alternative units to take the interaction into account. More precisely, Graph Convolution Network (GCN) and LSTM/GRU with different number of layers. For the experiment with GCN, we learned an adjacency matrix for every pair of persons and represented the interaction between them. We considered 4 GCN layers to obtain the refined poses. We also ablated the recurrent unit: GRU or LSTM [17]. Even though the MPJPE error of 4 LSTM layers is similar to that of 3 GRU layers, we considered the

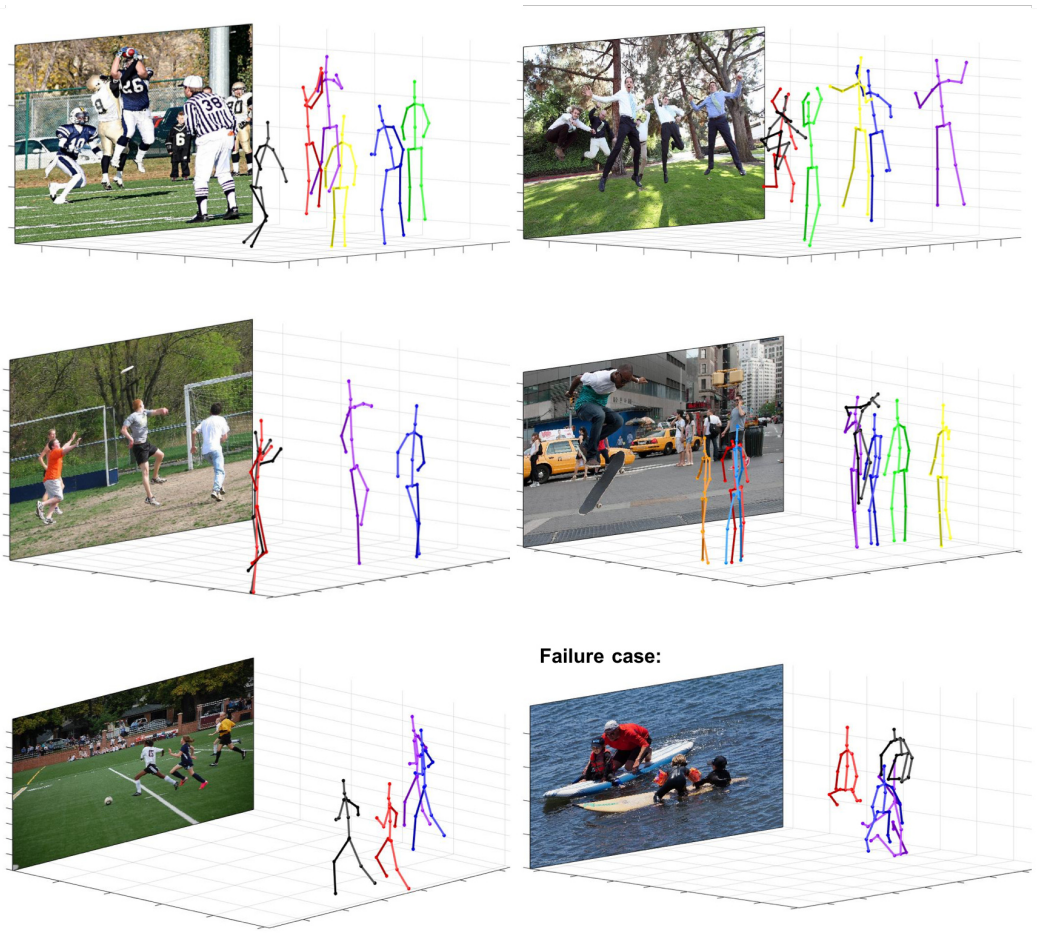

Figure 3: **Qualitative results on the COCO dataset**. For each pose, a darker color is used to represent the left side of the person. The bottom-right example corresponds to a failure case, as the 'red' and 'black' persons should be located in front of the scene, behind the 'blue' and 'purple' persons. This is caused by a misdetection on the root position of the input detected poses provided by the baseline network, while our network designed for refining the poses could not refine this kind of large deviation, because this large deviation caused by the baseline network hinder our PI-net from learning the correct context information for correctly interpreting and refining the prediction.

latter because it performs better after rigid alignment, and uses much less parameters which enables it to be trained more efficiently.

5. Conclusion

We propose PI-Net, a pose-interacting network that takes initial 3D body poses predicted by any pose estimator, and refine them leveraging on the mutual interaction that occurs in multi-person scenes. We learn such interactions using 3 main building blocks: a bi-directional RNN, a self-attention module, and a MLP. PI-Net is very flexible, lightweight and cost-efficient, and it could improve other approaches for multi-person 3D human pose estimation, establishing the new state-of-the-art. This line of work focuses on the interaction between people to improve perception results. In the future, we plan to extend this approach to reason on other contextual information such as objects or structures to better understand human actions and explore different ways to interpret relationships in the scene. Exploiting temporal priors [21, 46] and exploring other regression techniques such as robust deep regression [27] or regression adaptation [15, 22], are also other avenues we will explore.

6. Acknowledgement

We thank Yuming DU for inspiring discussions and feedback. This work has been partially funded by an Amazon Research Award and by the Spanish government under projects HuMoUR TIN2017-90086-R, Maria de Maeztu Seal of Excellence MDM-2016-0656, by the ANR JCJC ML3RI ANR-19-CE33-0008 and the ANR IDEX PIMPE ANR-15-IDEX-02.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2014.

[2] Ehud Barnea and Ohad Ben-Shahar. Exploring the bounds of the utility of context for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7412–7420, 2019.

[3] Abdallah Benzine, Florian Chabot, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Pandanet: Anchor-based single-shot multi-person 3d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6856–6865, 2020.

[4] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation from monocular images. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 584–588. IEEE, 2019.

[5] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *Proceedings of the European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[6] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *arXiv preprint arXiv:1812.08008*, 2018.

[7] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017.

[8] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7103–7112, 2018.

[9] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[10] Hai Ci, Chunyu Wang, Xiaoxuan Ma, and Yizhou Wang. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2262–2271, 2019.

[11] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[12] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Gregory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

[13] Rishabh Dabral, Nitesh B Gundavarapu, Rahul Mitra, Abhishek Sharma, Ganesh Ramakrishnan, and Arjun Jain. Multi-person 3d human pose estimation from monocular images. In *2019 International Conference on 3D Vision (3DV)*, pages 405–414. IEEE, 2019.

[14] Santosh K Divvala, Derek Hoiem, James H Hays, Alexei A Efros, and Martial Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1271–1278. IEEE, 2009.

[15] Diandra Fabre, Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Pierre Badin. Automatic animation of an articulatory tongue model from ultrasound images of the vocal tract. *Speech Communication*, 93:63–75, 2017.

[16] Mihai Fieraru, Mihai Zanfir, Elisabeta Oneata, Alin-Ionut Popa, Vlad Olaru, and Cristian Sminchisescu. Three-dimensional reconstruction of human interactions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7214–7223, 2020.

[17] Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. Learning to forget: Continual prediction with lstm. 1999.

[18] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, 2017.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[21] Alejandro Hernandez, Juergen Gall, and Francesc Moreno-Noguer. Human motion prediction via spatio-temporal inpainting. 2019.

[22] Thomas Hueber, Laurent Girin, Xavier Alameda-Pineda, and Gérard Bailly. Speaker-adaptive acoustic-articulatory inversion using cascaded gaussian mixture regression. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 23(12):2246–2259, 2015.

[23] 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.

[24] Wen Jiang, Nikos Kolotouros, Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Coherent reconstruction of multiple humans from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588, 2020.

[25] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[26] Laxman Kumarapu and Prerana Mukherjee. Animepose: Multi-person 3d pose estimation and animation. *arXiv preprint arXiv:2002.02792*, 2020.

[27] Stéphane Lathuilière, Pablo Mesejo, Xavier Alameda-Pineda, and Radu Horaud. Deepgum: Learning deep robust

regression with a gaussian-uniform mixture model. In *Proceedings of the European Conference on Computer Vision*, pages 202–217, 2018.

[28] Jiefeng Li, Can Wang, Wentao Liu, Chen Qian, and Cewu Lu. Hmor: Hierarchical multi-person ordinal relations for monocular multi-person 3d pose estimation. *arXiv preprint arXiv:2008.00206*, 2020.

[29] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision*, pages 332–347. Springer, 2014.

[30] Wenbo Li, Zhicheng Wang, Binyi Yin, Qixiang Peng, Yuming Du, Tianzi Xiao, Gang Yu, Hongtao Lu, Yichen Wei, and Jian Sun. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*, 2019.

[31] 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 *Proceedings of the European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[32] 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.

[33] 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 human pose estimation with a single rgb camera. *arXiv preprint arXiv:1907.00837*, 2019.

[34] 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.

[35] 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):1–14, 2017.

[36] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Camera distance-aware top-down approach for 3d multi-person pose estimation from a single rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10133–10142, 2019.

[37] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2823–2832, 2017.

[38] 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.

[39] Lourenço V Pato, Renato Negrinho, and Pedro MQ Aguiar. Seeing without looking: Contextual rescoring of object detections for ap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14610–14618, 2020.

[40] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10975–10985, 2019.

[41] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7025–7034, 2017.

[42] Leonid Pishchulin, Eldar Insafutdinov, Siyu Tang, Bjoern Andres, Mykhaylo Andriluka, Peter V Gehler, and Bernt Schiele. Deepcut: Joint subset partition and labeling for multi person pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4929–4937, 2016.

[43] Haibo Qiu, Chunyu Wang, Jingdong Wang, Naiyan Wang, and Wenjun Zeng. Cross view fusion for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4342–4351, 2019.

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

[45] Grégory 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.

[46] Edgar Simo-Serra, Carme Torras, and Francesc Moreno-Noguer. 3d human pose tracking priors using geodesic mixture models. 122(2):388–408, 2017.

[47] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[48] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2602–2611, 2017.

[49] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *Proceedings of the European Conference on Computer Vision*, pages 529–545, 2018.

[50] 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 *Proceedings of the European Conference on Computer Vision*, pages 601–617, 2018.

[51] Xiaolin K Wei and Jinxiang Chai. Modeling 3d human poses from uncalibrated monocular images. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1873–1880. IEEE, 2009.

[52] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. Dope: Distillation of part experts for whole-body 3d pose estimation in the wild. *arXiv preprint arXiv:2008.09457*, 2020.

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

[54] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5255–5264, 2018.

[55] Andrei Zanfir, Elisabeta Marinoiu, and Cristian Sminchisescu. Monocular 3d pose and shape estimation of multiple people in natural scenes-the importance of multiple scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2148–2157, 2018.

[56] Andrei Zanfir, Elisabeta Marinoiu, Mihai Zanfir, Alin-Ionut Popa, and Cristian Sminchisescu. Deep network for the integrated 3d sensing of multiple people in natural images. In *Advances in Neural Information Processing Systems*, pages 8410–8419, 2018.

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2881–2890, 2017.