PandaNet : Anchor-Based Single-Shot Multi-Person 3D Pose Estimation

Abdallah Benzine*,†, Florian Chabot*, Bertrand Luvison*, Quoc Cuong Pham*, Catherine Achard †
* CEA LIST Vision and Learning Lab for Scene Analysis
† Sorbonne University, CNRS, Institute for Intelligent Systems and Robotics

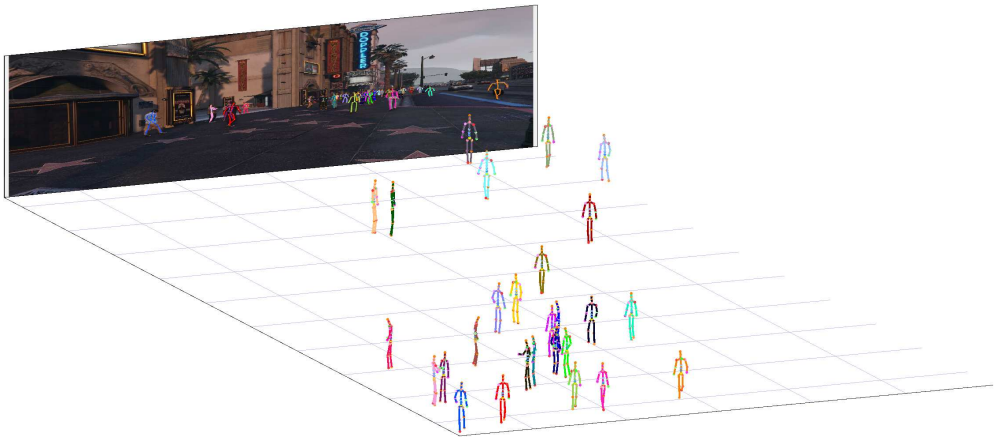


Figure 1: Qualitative results of PandaNet on JTA dataset [8] which consists in images with many people (up to 60), a large proportion of people at low resolution and many occlusion situations. Most of the previous 3D human pose estimation studies mainly focused on the single-person case or estimate 3D pose of few people at high resolution. In this paper, we propose an anchor-based and single-shot multi-person 3D pose estimation framework that allows the pose estimation of a large number of people at low resolution. Ground-truth translations and scales are used for visualisation.

Abstract

Recently, several deep learning models have been proposed for 3D human pose estimation. Nevertheless, most of these approaches only focus on the single-person case or estimate 3D pose of a few people at high resolution. Furthermore, many applications such as autonomous driving or crowd analysis require pose estimation of a large number of people possibly at low-resolution. In this work, we present PandaNet (Pose estimAtioN and Dectection Anchor-based Network), a new single-shot, anchor-based and multi-person 3D pose estimation approach. The proposed model performs bounding box detection and, for each detected person, 2D and 3D pose regression into a single forward pass. It does not need any post-processing to regroup joints since the network predicts a full 3D pose for each bounding box and allows the pose estimation of a possibly large number of people at low resolution. To manage people overlapping, we introduce a Pose-Aware Anchor Selection strategy. Moreover, as imbalance exists between different people sizes in the image, and joints coordinates have different uncertainties depending on these sizes, we propose a method to automatically optimize weights associated to different people scales and joints for efficient training. PandaNet surpasses previous single-shot methods on several challenging datasets: a multi-person urban virtual but very realistic dataset (JTA Dataset), and two real world 3D multi-person datasets (CMU Panoptic and MuPoTS-3D).

1. Introduction

3D human pose estimation is a common addressed problem in Computer Vision. It has many applications such as crowd analysis, autonomous driving, human computer interaction or motion capture. 3D pose is a low dimensional and interpretable representation that allows to understand and anticipate human behavior. Great progresses have been achieved thanks to large scale datasets with 2D annotations (LSP [13], MPII [1], COCO [22], CrowdPose [18]) and 3D annotations (Human 3.6M [12], MPI-INF-3DHP [25], MuCo-3D-HP [28], CMU Panoptic [14]). Nevertheless, this problem remains hard as the human body is an articu-

lated object whose terminal joints are very mobile and thus difficult to be precisely located. In addition, real-world applications require to handle a large number of people and crowded images like the ones in Figure 2b, 2c and 2d. To handle these challenging conditions, models need to be robust to people occlusions and to low resolution (*i.e* people that occupy a small portion of the image). They also need to be fast and to handle a large number of people. Most existing approaches focus on 3D pose estimation of either a single person or a limited number of people that are relatively close to the camera.

Although top-down and two-stage based methods are currently considered as best performing in the state of the art, these approaches become slow in crowded scenes as their computation complexity increases with the number of people. On the contrary, bottom-up approaches perform their forward pass with a constant complexity. Existing bottom-up single-shot methods rely on heatmaps prediction followed by complex post-processing steps to properly regroup joints detections into full human skeletons. 3D coordinates of joints are stored in maps at their corresponding 2D position. Consequently, if 2D localisation or 2D association of joints fails, 3D pose estimation will also fail. These approaches can also fail for other reasons. First, they lack precision at low resolution because of the downsampling factor between the input image and the predicted heatmaps. Second, heatmaps are usually not sharp enough to distinguish two very close joints of the same type. Finally, 3D coordinates of two overlapping joints cannot be stored at the same 2D location causing erroneous 3D pose estimation. For all these reasons, we believe that heatmap based approaches are not suited for robust 3D pose estimation for occluded people at low resolution.

In this paper, we introduce PandaNet (Pose estimAtioN and Dectection Anchor-based Network), a new single-shot approach that performs bounding box detection in a dense way and regresses 2D and 3D human poses for each detected person. To this end, three contributions are proposed.

First, an anchor based representation is adopted. An anchor that matches a subject stores its full 3D pose. This avoids problems induced by occlusion of joints. Additionally, this anchor-based formulation allows lower resolution outputs than heatmap one since a single output pixel is enough to store the entire subject's pose. This property is important to efficiently process people at low resolution.

Second, a Pose-Aware Anchor Selection strategy discards ambiguous anchors during inference. Indeed, ambiguous anchors overlap parts of several people and do not allow a readout of consistent 3D poses.

Third, an automatic weighting of losses with homoscedastic uncertainty handles imbalance between people sizes in the image and uncertainties associated to human pose predictions.

(a) (b) (c) (d)

Figure 2: **Different real-world contexts for 3D human pose estimation**. Recent 3D multi-person estimation approaches focus on 3D pose estimation of a few people close to the camera like in a). This context is yet challenging because of frequent inter-people occlusions. 3D pose estimation is even more difficult in applications such as sport analysis (b), autonomous driving (c) or crowd analysis (d) with a large number of people at low resolution.

Contrary to previous top-down multi-person approaches, PandaNet has a forward inference complexity that does not depend on the number of people in the image. It can efficiently process images with a large number of people (cf. Figure 1). The proposed model is validated on three 3D datasets. The first one is the Joint Track Auto dataset (JTA) [8], a rich, synthetic but very realistic urban dataset with a large number of people (up to 60) and occlusion situations. The second one is the Panoptic dataset [14], an indoor dataset with many interactions and social activities. The third one is the MuPoTS-3D dataset [28], a dataset with a reduced number of people but various indoor and outdoor contexts. Our results outperform those of previous single-shot methods on all these datasets.

2. Related Work

2D multi-person pose estimation. Two main approaches in multi-person 2D pose estimation cas be distinguished: top-down and bottom-up approaches. Methods of the former category first perform human detection then estimate a 2D pose within the detected bounding box. On the other hand, bottom-up approaches localise all human body keypoints in an input image and then group these keypoints into full 2D human skeletons. Each of them has advantages and disadvantages. Bottom-up methods are generally faster and seems more suited for crowded scenes since they process the entire image at once. However, available benchmarks show that top-down approaches are more accurate as they process all subjects at the same scale.

State of the art bottom-up approaches [4, 31, 16] differ on their association method. Cao *et al*. [4] propose Part Affinity Fields that are 2D vectors modeling the associations between children and parent joints. They are used to regroup 2D predictions of joints into full 2D skeletons. Newell *et al*. [31] perform this grouping by training the network to predict similar tag values to joints belonging to the

same person and different tag values for joints belonging to different people. Kreiss *el al.* [17] propose a bottom-up approach to handle people at low resolution in crowded images. Their model predicts Part Intensity Fields (PIF) that are similar to the offsets and heatmaps in [32] and Part Associative Field (PAF) that has a composite structure.

The methods in [11, 6, 32, 46, 19, 10] are top-down approaches. Mask R-CNN [11] detect keypoints as a segmentation task. The method in [32] performs 2D offsets and 2D heatmaps prediction and fuses these predictions to generate more precise heatmaps. Chen *et al*. [6] propose a cascaded pyramid network to generate 2D poses with a refinement process that focuses on hard keypoints. Xiao *et al*. [46] present a simple architecture with deep backbone and several upsampling layers.

While top-down approaches achieve higher 2D pose estimation scores in standard benchmarks than bottom-up approaches, most of these approaches fail in scenes with frequent and strong occlusions. Indeed, these methods depend on the predicted bounding boxes. In crowded scenes, bounding boxes, even if correct, may contain parts of other people. This situation is not well managed by existing methods. Li *et al*. [19] introduce a new benchmark to evaluate 2D human pose models on crowded scenes and a method that performs multi-peak predictions for each joint and a global maximum joint association. Golda *et al*. [10] propose an approach that explicitly detects occluded body parts, uses a data augmentation method to generate occlusions and exploits a synthetic generated dataset.

Single-person 3D pose estimation. There are two categories of single person 3D pose estimation approaches: direct and reconstruction approaches. Direct approaches estimate the 3D pose directly from an input image while reconstruction methods first take as input 2D poses provided by a 2D pose estimation model and lift them to the 3D space.

The approaches described in [9, 24] are reconstruction-based methods. Martinez *et al.* [24] regress 3D pose from 2D pose input by using a simple architecture with residual connections and batch normalisation. Fang *et al*. [9] use a pose grammar model that takes into account the connections between human joints. These reconstruction approaches are limited by the 2D pose estimator performance and do not take into account important images clues, such as contextual information, to make the prediction.

The models in [20, 34, 42, 43, 44, 50, 47] are direct approaches. Li *et al.* [20] simultaneously learn 3D pose regression and body part detection. Tekin *et al.* [44] predict 3D poses in an embedding space learned by an autoencoder. Pavlakos *et al.* [33] adopt a volumetric representation and a coarse to fine architecture to predict 3D poses. Sun *et al.* [42] take into account the connection structure between joints by proposing a compositional loss. Sun *et al.* [43] use the soft-argmax layer to extract 3D coordinates from a 3D volumetric representation in a differentiable way. Zhou *et al.* [50] use a geometric loss based on bones constraints to weakly supervise the depth regression module on *in the wild* images. Yang *et al.* [47] improve generalisation to *in the wild* images thanks to an adversarial loss.

Multi-person 3D pose estimation. Multi-person 3D pose estimation has been less studied. It is a difficult problem that adds to the 2D multi-person management difficulty, that of depth estimation. Zanfir *et al.* [49] estimate the 3D human shape from sequences of frames. A pipeline process is followed by a 3D pose refinement based on a non-linear optimisation process and semantic constraints. In a top-down approach, Rogez *et al.* [39, 40] generate human pose proposals that are classified into anchor-poses and further refined using a regressor. Moon *et al.* [29] propose a camera distance aware multi-person top-down approach that performs human detection (DetectNet), absolute 3D human localisation (RootNet) and root relative 3D human pose estimation (PoseNet). These approaches perform redundant estimations that need to be filtered or fused, and scales badly with a large number of people.

All existing single-shot methods estimate both 2D and 3D human poses and rely on heatmaps to detect individual joints in the image. Mehta *et al*. [28] propose a bottom-up approach system that predicts Occlusion-Robust Pose Maps (ORPM) and Part Affinity Fields [4] to manage multi-person 3D pose estimation even for occluded and cropped people. Benzine *et al.* [2, 3] perform single-shot multi-person 3D pose estimation by extending the 2D multi-person model in [31] to predict ORPM. ORPM based methods predict a fixed number of 2D heatmaps and ORPM, whatever the number of people in the image. 3D coordinates are stored multiple times in the ORPM allowing the readout of 3D coordinates at non occluded and reliable 2D positions. Nevertheless, this formulation implies potential conflicts when similar joints of different people overlap. In the same way, MubyNet [48] also uses a fixed number of output maps to store 2D and 3D poses of all people in the image. However, the full 3D pose vector is stored at all 2D positions of the subject skeleton increasing the number of potential conflicts. The model learns to score the possible associations of joints to limbs and a global optimisation problem is solved to group the joints into full skeletons. XNect [27] estimates 3D poses in two steps. The first step improves the method of [28] by encoding only the joints' immediate context, which reduces the number of potential conflicts. The second step refines the 3D poses.

PandaNet is a single-shot approach like [48, 28, 2, 27] but is anchor-based rather than heatmap-based. It is based on LapNet [5], a single-shot object detection model which has today the best accuracy/inference time trade-off. Unlike LapNet that is a 2D object detector, PandaNet is intended for multi- person 3D pose estimation and differs from Lap-

Net in the prediction heads (introduced in subsection 3.1), in the anchor selection strategy (described in subsection 3.4) and in the automatic weighting of losses (described in subsection 3.5). It efficiently processes images with many people, strong occlusion and various sizes in the image with a complexity that does not depend on their number.

3. Method

3.1. Overview

Given an input image, PandaNet predicts a dense set of human bounding boxes with their associated confidence scores, 2D and 3D poses. These boxes are then filtered using non-maximum suppression to provide the final detections and human poses. As in most detection approaches [37, 23, 35, 21, 5], our model uses predefined anchors. These anchors are computed on the training dataset with the clustering method used in [5, 35]. We define N_A to be the number of predefined human anchors used in the model, N_K the number of human joints and H and W the height and the width of the network output. The model returns: (1) Score maps $\hat{C} \in \mathbb{R}^{H \times W \times N_A}$ that contain the probability of an anchor to contain a subject (2) Box offsets maps $\hat{B} \in \mathbb{R}^{H \times W \times N_A \times 4}$ (3) 2D joints coordinates maps $\hat{P}^{2D} \in \mathbb{R}^{H \times W \times N_A \times N_K \times 2}$ that contain the full 2D pose vectors expressed relatively to their corresponding anchor (4) 3D joints coordinates maps $\hat{P}^{3D} \in \mathbb{R}^{H \times W \times N_A \times N_K \times 3}$ that contain root relative 3D human poses.

PandaNet is a multi-task network based on LapNet, the single-shot object detection model proposed in [5] which has today the best accuracy/inference time trade-off. The architecture of the proposed model, detailed in Figure 3, slightly differs from LapNet. First, sub-pixel convolutions are applied [41] to the feature maps to obtain higher resolution maps that are crucial to detect and estimate the human pose of people at low resolution. Secondly, a 2D pose and 3D pose regression heads are added.

3.2. Anchor-based Multi-person Pose Formulation

For a given image I, we define $\mathcal{B} = \{b_n \in \mathbb{R}^4\}$ as the set of ground truth bounding boxes $n \in [1, \dots, N]$ and N is the number of visible people. $\mathcal{P}^{2D} = \{p_n^{2D} \in \mathbb{R}^{2 \times N_K}\}$ and $\mathcal{P}^{3D} = \{p_n^{3D} \in \mathbb{R}^{3 \times N_K}\}$ are the sets of corresponding 2D and 3D human poses.

In order to train PandaNet, a grid of anchors $A \in \mathbb{R}^{H \times W \times N_A \times 4}$ is defined. $A_{i,j,a}$ is an element of this grid at output position (i, j) for anchor a. Let $B \in \mathbb{R}^{H \times W \times N_A \times 4}$ be the grid of matched bounding boxes, each of its element is defined as:

$$B_{i,j,a} = \underset{b_n \in \mathcal{B}}{\operatorname{argmax}}\, IoU(b_n, A_{i,j,a}) \tag{1}$$

Similarly, P^{2D} and P^{3D} are defined respectively as the grids of matched 2D poses and 3D poses:

$$P^{2D}_{i,j,a} = p_n^{2D} \mid b_n = B_{i,j,a} \tag{2}$$

$$P^{3D}_{i,j,a} = p_n^{3D} \mid b_n = B_{i,j,a} \tag{3}$$

In other words, $P^{2D}_{i,j,a}$ and $P^{3D}_{i,j,a}$ are the 2D and 3D human poses of the subject matched by the anchor $A_{i,j,a}$

The Per-Object Normalised Overlap (PONO) map [5] O is used. $O_{i,j,a}$ is the IoU between anchor $A_{i,j,a}$ and ground truth $B_{i,j,a}$, normalised by the maximum overlap between $B_{i,j,a}$ and all matched anchors to this ground truth.

The positive anchors $\mathcal{A}^+$ are the set of matched anchors that have a PONO value greater than 0.5. Only bounding boxes and human poses associated to anchors in $\mathcal{A}^+$ will be supervised, like described in the next subsection.

3.3. Bounding box offsets and human poses supervision

3.3.1 IoU based bounding-box offsets supervision

Most detection approaches use SmoothL1 or Mean Squared Error losses to predict bounding box offsets that fit the ground truth bounding boxes. More recently, some methods prefer to optimize the Intersection over Union (IoU) loss [45, 5] or its extension [38] taking benefit of its invariance to scale. We also used the IoU loss to supervise bounding box offsets prediction and, for an anchor a at location (i, j), we define the overlap function as:

$$\hat{O}_{i,j,a} = IoU(B_{i,j,a}, \hat{B}_{i,j,a}) \tag{4}$$

where $\hat{B}_{i,j,a}$ is the predicted box, *i.e* the anchor $A_{i,j,a}$ transformed with estimated offsets. The pixel-wise localisation loss is then obtained with:

$$\mathcal{L}_{loc}(i,j,a) = \begin{cases} \left\| 1 - \hat{O}_{i,j,a} \right\|^2, & \text{if } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

3.3.2 IoU based 2D human pose supervision

While our main objective is single-shot multi-person 3D pose estimation, PandaNet also regresses 2D human poses for two reasons. Firstly, the predicted 2D poses are needed in the pose-aware pixel-wise classification loss defined in subsection 3.4. Secondly, by minimizing the reprojection loss between a 2D human pose and a root relative 3D human pose, one can obtain the 3D human pose in the camera reference. Regressing 2D human poses is challenging because of large variations in scale between people. So, we introduce a IoU loss to supervise this step. We designate by $P^{2D}_{i,j,a,k}$ the ground-truth 2D coordinates in the anchor coordinate system of the joint k of the subject matched with the

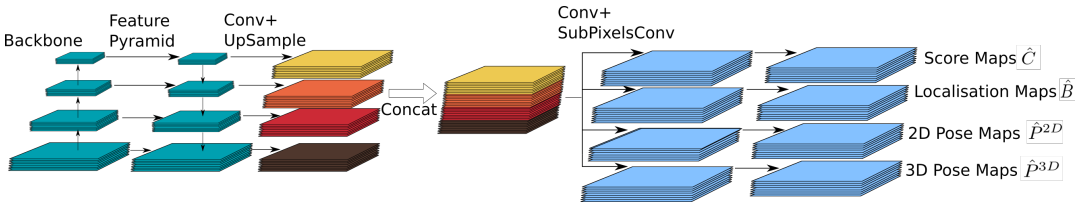


Figure 3: PandaNet architecture. The input image is passed through a backbone network. A second stage is used to compute pyramid feature maps at several resolutions and semantic levels (like done by FPN [21]). Four 3x3 convolutions are applied to these feature maps. The resulting maps are then upsampled to the size of the highest resolution feature map. After multi-scale feature concatenation and subpixels convolution, four convolutional heads are used to provide the four outputs. Each head is composed by four 3x3 convolutions and one final convolutional layer for the output.

anchor a. These coordinates are obtained from the coordinates in the image space by translating them to the center of the anchor and dividing them by the width and the height of the anchor. $\hat{P}^{2D}_{i,j,a,k}$ are the corresponding predicted coordinates in the anchor space. Two unit squares in the anchor space, $\hat{S}_{i,j,a,k}$ and $S_{i,j,a,k}$, centred at positions $\hat{P}^{2D}_{i,j,a,k}$ and $P^{2D}_{i,j,a,k}$ are defined to compute the IoU loss and the pixel-wise 2D pose loss for joint k:

$$\hat{O}^{2D}_{i,j,a,k} = IoU(S_{i,j,a,k}, \hat{S}_{i,j,a,k}) \tag{6}$$

$$\mathcal{L}_{2D}(i,j,a,k) = \begin{cases} \left\| 1 - \hat{O}^{2D}_{i,j,a,k} \right\|^2, & \text{if } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

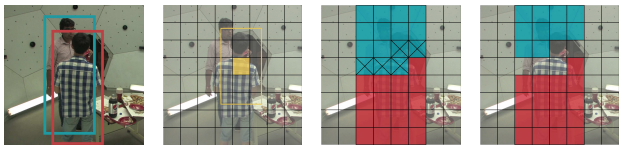
3.3.3 3D human pose supervision

PandaNet is trained to predict scale normalised 3D human poses translated to the pelvis. The sum of the subject bones length is equal to 1. As all 3D poses are predicted at the same scale, an Euclidean distance is used as supervision. The pixel-wise 3D pose loss for joint k between the ground-truth 3D joints coordinates $P^{3D}_{i,j,a,k}$ and their corresponding predicted coordinates $\hat{P}^{3D}_{i,j,a,k}$ is defined by :

$$\mathcal{L}_{3D}(i,j,a,k) = \begin{cases} \left\| P^{3D}_{i,j,a,k} - \hat{P}^{3D}_{i,j,a,k} \right\|^2, & \text{if } A_{i,j,a} \in \mathcal{A}^+ \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

3.4. Pose-Aware Anchor Selection

As illustrated in Figure 4, some of the positive anchors in $\mathcal{A}^+$ are not suited for the readout of consistent human poses. When several subjects overlap, these anchors may contain more than one person. This can lead to erroneous predicted bounding boxes and incorrect human poses. Consequently, at inference, precise readout locations are needed to determine the final bounding boxes and poses. To do so, the network should be trained to consider ambiguous anchors as negative. We define $\mathcal{A}^{++}$ to be the set of non ambiguous readout anchors.



(a) Input Image (b) Grid of Anchors $\mathcal{A}$. (c) Matched Anchors $\mathcal{A}^+$ (d) Selected Anchors $\mathcal{A}^{++}$

Figure 4: **Pose-Aware Anchors Selection**. A grid of anchors $\mathcal{A}$ is first computed at all output 2D positions (Figure 4b). An example of an anchor is depicted in yellow. The matched anchors $\mathcal{A}^+$ correspond to anchors with a sufficient PONO (depicted in red and blue in Figure 4c). Nevertheless, some of these anchors are ambiguous (crossed anchors in 4c) as they correspond to overlapping persons. They are filtered by the Pose-Aware Anchors Selection strategy to obtain the set of non-ambiguous positive readout anchors $\mathcal{A}^{++}$ depicted in Figure 4d (best viewed in color).

A way to filter $\mathcal{A}^+$ to get $\mathcal{A}^{++}$ is to threshold the product of the overlap between ground truth and predicted bounding boxes $\hat{O}_{i,j,a}$ and the PONO value $O_{i,j,a}$, as it is done in [5]. In other words, an anchor belongs to $\mathcal{A}^{++}$ if the box predicted by this anchor correctly fit its associated ground truth. For detection purpose this strategy may be sufficient to solve ambiguities, but for pose estimation such a filtering is too coarse. Anchors in $\mathcal{A}^{++}$ must lead to the readout of valid and unambiguous human poses. To this end, we introduce a Pose-Aware Anchor Selection strategy based on 2D poses overlap. This overlap $\hat{O}^{2D}_{i,j,a}$ is defined as the mean of $\hat{O}^{2D}_{i,j,a,k}$ for all joints k of the subject.

Thus, the Positive Readout Anchors Labels ($C_{i,j,a}$) are defined by :

$$C_{i,j,a} = \begin{cases} 1 & \text{if } O_{i,j,a} \times \hat{O}^{2D}_{i,j,a} > 0.5 \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

The pixel-wise classification loss is then defined by:

$$\mathcal{L}_{cls}(i,j,a) = \mathcal{H}(C_{i,j,a}, \hat{C}_{i,j,a}) \tag{10}$$

where $\mathcal{H}$ is the standard binary cross-entropy.

3.5. Automatic weighting of losses with homoscedastic uncertainty

A classical problem in training a multi-task network is to properly weight each task. Kendall *et al.* [15] propose a loss based on the homoscedastic uncertainty (*i.e* independent of the input data) to weight multiple losses in a multi-task network. Another issue in single-shot multi-person pose estimation is the imbalance between subject's sizes in the image. In real-world images, there are both a large range of distances to the camera and an imbalance in people sizes in the image. In Lapnet [5], anchor weights are learned to solve this problem for rigid object detection. In multi-person pose estimation tasks of PandaNet, uncertainties related to joints have to be managed. Indeed, as joints have different degrees of freedom, predictions associated to hips are more certain than predictions associated to hands for instance. Uncertainty also depends on people sizes in the image. A network is less precise for people at low resolution than for high resolution people. Furthermore, far from the camera people are more prone to occlusions than other people making the regressed coordinates associated to these people more uncertain. This is why we propose to learn joint specific regression weights for each predefined anchor and introduce the following loss functions:

$$\begin{aligned}\mathcal{L}_{cls} = &\frac{\lambda_{cls}}{HWN_A}\sum_a \lambda^a_{cls}\sum_{i,j}\mathcal{L}_{cls}(i,j,a) \\ &+ \log(\frac{1}{\lambda_{cls}}) + \frac{1}{N_A}\sum_a \log(\frac{1}{\lambda^a_{cls}})\end{aligned} \tag{11}$$

$$\begin{aligned}\mathcal{L}_{loc} = &\frac{\lambda_{loc}}{N^+}\sum_a \lambda^a_{loc}\sum_{i,j}\mathcal{L}_{loc}(i,j,a) \\ &+ \log(\frac{1}{\lambda_{loc}}) + \frac{1}{N_A}\sum_a \log(\frac{1}{\lambda^a_{loc}})\end{aligned} \tag{12}$$

$$\begin{aligned}\mathcal{L}_{2D} = &\frac{\lambda_{2D}}{N_K N^+}\sum_{i,j,a,k}\lambda^{a,k}_{2D}\mathcal{L}_{2D}(i,j,a,k) \\ &+ \frac{1}{N_K N_A}\sum_{a,k}\log(\frac{1}{\lambda^{k,a}_{2D}})\end{aligned} \tag{13}$$

$$\begin{aligned}\mathcal{L}_{3D} = &\frac{\lambda_{3D}}{N_K N^+}\sum_{i,j,a,k}\lambda^{a,k}_{3D}\mathcal{L}_{3D}(i,j,a,k) \\ &+ \frac{1}{N_K N_A}\sum_{a,k}\log(\frac{1}{\lambda^{k,a}_{3D}})\end{aligned} \tag{14}$$

where λ_{cls}, λ_{loc}, λ_{Pose2D} and λ_{Pose3D} are the task weights, λ^a_{cls} and λ^a_{loc} are the anchors weights and $\lambda^{a,k}_{Pose2D}$ and $\lambda^{a,k}_{Pose3D}$ are the anchor-joint regression weights. N^+ is the number of anchors in A^+. All weights λ are trainable variables. All terms $\log(\frac{1}{\lambda})$ are regularisation terms that avoid all λ to converge to 0. The final total loss is :

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{loc} + \mathcal{L}_{2D} + \mathcal{L}_{3D} \tag{15}$$

4. Experimental Results

PandaNet's performance is evaluated on three datasets: JTA [8], CMU-Panoptic [14] and MuPoTS-3D [28].

Evaluation Metrics: To evaluate multi-person 3D pose approaches, we use two metrics. The first one is the Mean per Joint Position Error (MPJPE) that corresponds to the mean Euclidean distance between the ground truth and the prediction for all people and all joints. The second one is the 3DPCK which is the 3D extension of the Percentage of Correct Keypoints (PCK) metric used for 2D pose evaluation. A joint is considered correctly estimated if the error in its estimation is less than 150mm. If an annotated subject is not detected by an approach, we consider all of its joints to be incorrect in the 3DPCK metric. The human detection performance is evaluated with the Average Precision (AP) used in the PASCAL VOC challenge [7].

Training Procedure: The method was implemented and tested with TensorFlow 1.12. In all our experiments, the model is trained with mini-batches of 24 images. SGD optimiser is used with a momentum of 0.9, an initial learning rate of 0.005 and a polynomial decay policy of 0.9. Random crops and random scales are used for data augmentation. Synchronized batch normalisation across GPU is used. Darknet-53 [36] is used as backbone. The number of anchors N_A is set to 10 in all experiments.

4.1. JTA dataset results

JTA (Joint Track Auto) is a dataset for human pose estimation and tracking in urban environment. It was collected from the realistic video-game the Grand Theft Auto V and contains 512 HD videos of 30 seconds recorded at 30 fps. The collected videos feature a vast number of different body poses, in several urban scenarios at varying illumination conditions and viewpoints. People perform different actions like walking, sitting, running, chatting, talking on the phone, drinking or smoking. Each image contains a number of people ranging between 0 and 60 with an average of 21 people. The distance from the camera ranges between 0.1 to 100 meters, resulting in pedestrian heights between 20 and 1100 pixels. No existing dataset with annotated 3D poses is comparable with JTA dataset in terms of number and sizes of people. An input image size of 928x576 is used.

Anchor Selection	AP	3DPCK
No	84.1	80.7
BB-Aware [5]	85.1	81.9
Pose-Aware (Ours)	**85.3**	**83.2**

Table 1: Influence of the Anchor Selection Strategy. All the models are trained with the Automatic Weighting of Losses.

task	anchor	joint	AP	3DPCK
1	1	1	21.7	15.8
learned	1	1	84.1	80.8
learned	learned	1	85.2	81.7
learned	learned	learned	**85.3**	**83.2**

Table 2: Influence of the Automatic Weighting of Losses. task, anchor and joint represent the type of trainable λ weights. All the models are trained with the Pose-Aware Anchor Selection Strategy.

4.1.1 Ablation Studies

Pose-Aware Anchor Selection strategy: Table 1 results show the effectiveness of the Pose-Aware Anchor Selection. We compare three variants of PandaNet. The first variant (first line) is a model where no anchor selection strategy is used. It corresponds to a model where only the PONO overlap $O_{i,j,a}$ is considered in equation 9. Using the Bounding-box Aware Anchor Selection [5] (second row), improves the model performance over this baseline. Box detection and 3D pose estimation take all benefit of this anchor selection strategy. Using the proposed Pose-Aware Anchor Selection (third row) maintain the AP value while improving the 3DPCK, showing its effectiveness for choosing better anchors for 3D pose estimation.

Automatic Weighting of Losses: The influence of Automatic Weighting of Losses is detailed in Table 2. When the λ's are all set to 1 (first line) and not trained, the model has poor performance on all tasks. Learning task-specific λ_{loc}, λ_{cls}, λ_{2D} and λ_{3D} (second row) allow the network to converge and to achieve good performances on all tasks. Learning anchor weights λ^a_{loc} and λ^a_{cls} (third row) improves detection and 3D pose estimation performances. The best results are obtained when all λ's are learned, showing the importance of the proposed automatic weighting of losses.

4.1.2 Comparison with prior work

The approach in [3] is the only method that provides 3D results on the JTA dataset. We compare PandaNet with the best model in [3] *i.e* the model with multi-scale inference.

Table 3 provides 3DPCK results according to the distance of people to the camera. PandaNet outperforms the model of Benzine *et al*. [3] on all camera distances demonstrating the ability of PandaNet to properly process people at all scales. While our model achieve very good results for people close to the camera (less than 20m), it also correctly handles people who are further from the camera.

Dist.	<10	10-20	20-30	30-40	>40	All
[3]	55.8	61.6	42.2	36.0	41.7	43.9
Ours	**95.6**	**93.7**	**87.3**	**80.5**	**71.2**	**83.2**

Table 3: Distance wise 3DPCK on the JTA dataset. Distance are in meters.

Table 4 provides joint-wise results on the JTA dataset. PandaNet outperforms the model of Benzine *et al*. [3] for all joints. In particular, it has no difficulties to estimate 3D coordinates for the joints that have the fewest degrees of freedom (head, neck, clavicles , hips and spines) with a 3DPCK for these joints greater than 92%. PandaNet increases the 3DPCK for the shoulders by 44.6% and for the elbows by 34.9%. Terminal joints (wrists and ankles) are the most difficult joints with a 3DPCK of 60.1% and 58.0% for these joints against 19.0% and 8.9% for [3].

4.2. CMU-Panoptic results

CMU Panoptic [14] is a dataset containing images with several people performing different scenarios in a dome where many cameras are placed. Although it was acquired in a simple and controlled environment, this dataset is challenging because of complex interactions and difficult camera viewpoints. We evaluate PandaNet using the protocol used in [48, 49, 2]. The test set is composed of 9600 frames from HD cameras 16 and 30 and for 4 scenarios: Haggling, Mafia, Ultimatum, Pizza. The model is trained on the other 28 HD cameras of CMU Panoptic. An input image size of 512x320 is used on all Panoptic experiments.

On this dataset, PandaNet improves the results over the recent state of the art methods on all scenarios (Table 5). The average MPJPE is improved by 25.8mm compared to the best previous approach. While the results on JTA prove the ability of the model to deal with realistic urban scenes with many people at low resolution, results on the Panoptic dataset show that the approach is effective to manage people overlaps and crops that frequently occur in this dataset.

4.3. MuPoTS-3D results

MuPoTS-3D [28] is a dataset containing 20 sequences with ground truth 3D poses for up to three subjects. PandaNet is trained on the MuCo-3DHP dataset that is generated by compositing the existing MPI-INF-3DHP 3D single-person pose estimation dataset [26], and on the COCO-dataset [22] to ensure better generalisation. Each mini-batch consists of half MuCo-3DHP and half COCO images. For COCO data, the loss value for the 3D regres-

Method	head	neck	clavicles	shoulders	elbows	wrists	spines	hips.	knees	ankles	all
[3]	41.1	44.6	44.9	33.8	27.2	19.0	74.4	73.9	25.7	8.9	43.9
Ours	**92.7**	**99.1**	**97.0**	**78.4**	**72.1**	**60.1**	**99.9**	**87.8**	**71.8**	**58.0**	**83.2**

Table 4: Joint wise 3DPCK.

Method	Haggling	Mafia	Ultimatum	Pizza	Mean
[48]	140.0	165.9	150.7	156.0	153.4
[49]	72.4	78.8	66.8	94.3	72.1
[2]	70.1	66.6	55.6	78.4	68.5
Ours	**40.6**	**37.6**	**31.3**	**55.8**	**42.7**

Table 5: MPJPE in mm on the Panoptic dataset.

Method		3DPCK
Two-Stage	LCR-Net[39]	53.8
	LCR-Net++ [40]	70.6
	Moon *et al*. [29]	**81.8**
Single-Shot	Mehta *et al*. [28]	66.0
	XNect [27]	70.4
	PandaNet	**72.0**

Table 6: 3DPCK on the MuPoTS-3D dataset.

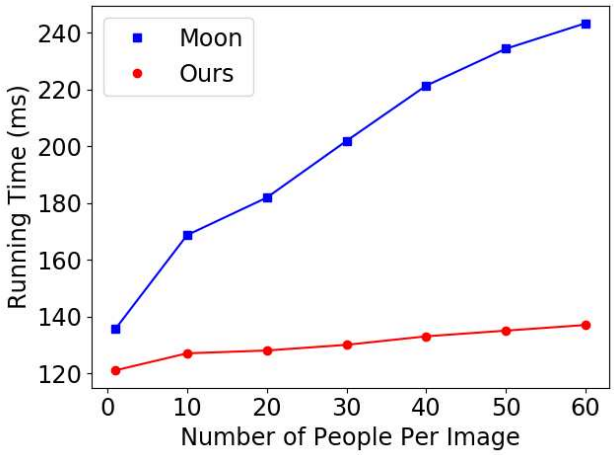


Figure 5: Running time comparison with the approach of Moon *et al*. [30] according to the number of people per image. Experiments are performed on a NVIDIA Titan X GPU. Images come from JTA Dataset [8].

sion task is set to zero. An input image size size of 512x512 is used and Subpixel convolutions are removed.

Table 6 provides 3DPCK results on this dataset. PandaNet achieve higher 3DPCK than previous single-shot approaches. It improves over an ORPM method [28] by 6% and over XNect [27] by 1.6%. XNect is composed of two different models. The first one predicts partial 2D and 3D pose encoding and the second one refines these encodings to get final full 3D poses. Consequently, the weaknesses of the first model (like joints occlusions and people crops) are compensated by the second one. We achieve better results with a single model without any refinement process. Compared to two-stage models, PandaNet achieves better results than LCR-Net [39] and LCR-Net++ [40]. Compared to the approach of Moon *et al*. [30], PandaNet has a lower 3DPCK. This top-down approach uses an external two-stage object detector (Faster-R CNN [37]) to compute human bounding boxes and forward each cropped subject to a single-person 3D person approach [43]. Therefore, the computation complexity of their model depends on the number of people in the image like illustrated in Figure 5. If the number of people is large, this approach scales badly. On the contrary, the proposed single-shot model allows a nearly constant inference time regarding the number of people. The inference time of PandaNet is about 140ms for images with 60 people on a NVIDIA Titan X GPU.

5. Conclusion

PandaNet is a new anchor-based single-shot multi-person pose estimation model that efficiently handles scene with a large number of people, large variation in scale and people overlaps. This model predicts in a single-shot way people bounding boxes and their corresponding 2D and 3D pose. To properly manage people overlaps, we introduce a Pose-Aware Anchor Selection strategy that discards ambiguous anchors. Moreover, an automatic weighting has been provided for three main purposes. It balances task-specific losses, it compensates imbalance in people sizes and it manages uncertainty related to joints coordinates.

The experiments validate the proposed Anchor-based Multi-person Pose Regression framework and prove the importance of the Pose-Aware Anchor Selection strategy and of the Automatic Weighting. Furthermore, large-scale experiments, on JTA, CMU Panoptic, and MuPoTS-3D datasets demonstrate that PandaNet outperforms previous single-shot state of the art methods.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, June 2014. 1

[2] 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 *ICIP*, 2019. 3, 7, 8

[3] Abdallah Benzine, Bertrand Luvison, Quoc Cuong Pham, and Catherine Achard. Deep, robust and single shot 3d multi-person human pose estimation in complex images. *arXiv preprint arXiv:1911.03391*, 2019. 3, 7, 8

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *CVPR*, 2017. 2, 3

[5] Florian Chabot, Mohamed Chaouch, and Quoc Cuong Pham. Lapnet : Automatic balanced loss and optimal assignment for real-time dense object detection. *arXiv preprint arXiv:1911.01149*, 2019. 3, 4, 5, 6, 7

[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *CVPR*, 2018. 3

[7] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *ICCV*, 2015. 6

[8] 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. *ECCV*, 2018. 1, 2, 6, 8

[9] Hao-Shu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. *AAAI Conference on Artificial Intelligence*, 2018. 3

[10] Thomas Golda, Tobias Kalb, Arne Schumann, and Jüergen Beyerer. Human Pose Estimation for Real-World Crowded Scenarios. In *AVSS*, 2019. 3

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[12] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *PAMI*, 36(7), 2014. 1

[13] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *BMVC*, 2010. 1

[14] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Godisart, Bart Nabbe, Iain Matthews, et al. Panoptic studio: A massively multiview system for social interaction capture. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 1, 2, 6, 7

[15] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, 2018. 6

[16] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Multiposenet: Fast multi-person pose estimation using pose residual network. In *ECCV*, 2018. 2

[17] Sven Kreiss, Lorenzo Bertoni, and Alexandre Alahi. Pifpaf: Composite fields for human pose estimation. In *CVPR*, 2019. 3

[18] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. 2018. 1

[19] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. In *CVPR*, 2019. 3

[20] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 3

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4, 5

[22] 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 *ECCV*, 2014. 1, 7

[23] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 4

[24] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 3

[25] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision*, 2017. 1

[26] Dushyant Mehta, Helge Rhodin, Dan Casas, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation using transfer learning and improved cnn supervision. arxiv preprint. *arXiv preprint arXiv:1611.09813*, 2016. 7

[27] 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. 3, 8

[28] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Srinath Sridhar, Gerard Pons-Moll, and Christian Theobalt. Single-shot multi-person 3d body pose estimation from monocular rgb input. *3DV*, 2017. 1, 2, 3, 6, 7, 8

[29] 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. *ICCV*, 2019. 3, 8

[30] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. Multi-scale aggregation r-cnn for 2d multi-person pose estimation. *CVPR*, 2019. 8

[31] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *NIPS*, 2017. 2, 3

[32] George Papandreou, Tyler Zhu, Nori Kanazawa, Alexander Toshev, Jonathan Tompson, Chris Bregler, and Kevin Murphy. Towards accurate multi-person pose estimation in the wild. In *CVPR*, 2017. 3

[33] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *ECCV*, 2016. 3

[34] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *CVPR*, 2017. 3

[35] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 4

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6

[37] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 4, 8

[38] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 4

[39] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR*, 2017. 3, 8

[40] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net++: Multi-person 2d and 3d pose detection in natural images. *TPAMI*, 2019. 3, 8

[41] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *CVPR*, 2016. 4

[42] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 3

[43] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 3, 8

[44] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. *arXiv preprint arXiv:1605.05180*, 2016. 3

[45] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. *arXiv preprint arXiv:1904.01355*, 2019. 4

[46] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 3

[47] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 3

[48] 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 *CVPR*, 2018. 3, 7, 8

[49] 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 *NIPS*, 2018. 3, 7, 8

[50] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 3