Self-supervised 3D Human Pose Estimation from a Single Image

Anonymous CVPR submission

Paper ID 6

Abstract

We propose a new self-supervised method for predicting 3D human body pose from a single image. The prediction network is trained from a dataset of unlabelled images depicting people in typical poses and a set of unpaired 2D poses. By minimising the need for annotated data, the method has the potential for rapid application to pose estimation of other articulated structures (e.g. animals). The self-supervision comes from an earlier idea exploiting consistency between predicted pose under 3D rotation. Our method is a substantial advance on state-of-the-art self-supervised methods in training a mapping directly from images, without limb articulation constraints or any 3D empirical pose prior. We compare performance with state-of-the-art self-supervised methods using benchmark datasets that provide images and ground-truth 3D pose (Human3.6M, MPI-INF-3DHP). Despite the reduced requirement for annotated data, we show that the method outperforms on Human3.6M and matches performance on MPI-INF-3DHP. Qualitative results on a dataset of human hands show the potential for rapidly learning to predict 3D pose for articulated structures other than the human body.

1. Introduction

Estimating 3D pose for articulated objects is a long-standing problem. Its foundations arise from the early days of computer vision with model-based approaches representing the human body as an articulated structure of parts [10, 26]. Interest in estimating 3D human pose grew within the computer vision community motivated by the many real-world applications, for example, pedestrian detection [16], human-computer interaction [42], video surveillance [43], and sports analysis [30].

Initial work on estimating 3D pose addressed this problem by extracting a set of hand-crafted features, for example, segmentation masks [1]. Other early approaches, such as exemplar-based methods, use extensive datasets of 3D poses (commonly constructed from motion capture data) to search for the optimal 3D pose given its 2D projection

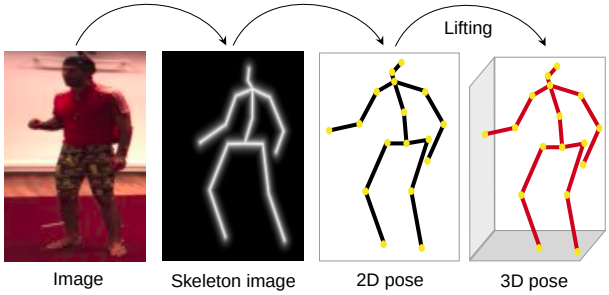

Figure 1. **3D pose estimation pipeline.** Our approach jointly learns to estimate 3D pose from an image via intermediate representations of 2D pose. The pipeline is embedded within a larger network for end-to-end training.

[2, 9, 15]. Subsequently, these approaches were surpassed in performance by deep learning methods, initially using supervised learning to regress from images to joint positions [37] or heatmaps [27, 36]. While annotations for 2D joint positions in the image plane are relatively easy to obtain, getting ground truth 3D joint positions from images alone is not straightforward.

Because of the limited availability of 2D and 3D pose annotations, an immense amount of available data in the form of images remains unexploited for training in pose estimation. Although there are many annotated datasets for human pose estimation, the situation is very different for non-human articulated structures such as animals. Recently, unsupervised and self-supervised methods for 3D human pose estimation have made progress using this unlabelled data (at least lacking 3D annotations) and have demonstrated that it is possible to learn to estimate 3D poses from 2D poses relying on rotational consistency [3, 5, 41], multi-view setup [13, 31], and 3D structure constraints e.g. joint angles between limbs [18]. However, what will happen if we do not assume the availability of paired 2D poses? Is it still possible to train a model to predict 3D poses? What are the minimum assumptions to make this possible?

We have proposed a method which, for the first time, learns to map between images and 3D pose without requiring 3D pose annotations or paired 2D pose annotations.

Training only needs a set of unlabelled images depicting people in different poses and an unrelated set of 2D human poses. The motivation is twofold: (1) there is the potential for exceeding current levels of performance by training on massive unlabelled datasets, and (2) the method could, in principle, be applied to articulated structures (e.g. animals) where little or no 2D/3D annotated data is available. In our proposed method, we learn a 2D pose predictor and a 3D 'lifting' function to produce 3D joint positions from unlabelled images (summarised in Figure 1) in an end-to-end learning framework.

Our method simultaneously learns 2D and 3D pose representations in a largely unsupervised fashion, requiring only an empirical prior on unpaired 2D poses. We demonstrate its effectiveness on Human3.6M [11] and MPI-INF-3DHP [23] datasets, two of the most popular benchmarks for human pose estimation. We also show the method's adaptability to other articulated structures using a synthetic dataset of human hands [33]. In experiments, the approach outperforms state-of-the-art self-supervised methods that estimate 3D pose from images and require higher supervision in training. Overall, our method has the following advantages:

- It does not assume any 3D pose annotations or paired 2D pose annotations.
- It holds the potential for quickly adapting to 3D pose prediction for other articulated structures (e.g. animals and jointed inanimate objects).

2. Related Work

Our method broadly relates to prior work that estimates 3D human pose directly from images, and mainly to self-supervised deep learning methods. However, it also draws inspiration from earlier work on the estimation of 3D pose from 2D pose. Therefore, we review both perspectives, regardless of the degree of supervision required for training.

3D pose from 2D pose

A range of methods take as input 2D poses and lift them to 3D space. Frequently, the 2D poses come from an off-the-shelf 2D pose estimator, or they are simply annotations for a given dataset. Early techniques for estimating 3D poses from 2D joint positions rely on classical classification algorithms and physical constraints. For example, given the joint connectivity and bone lengths, [19] use binary decision trees to estimate the two possible states of a joint with respect to its parent. Other methods implement the nearest neighbour algorithm with large datasets of 3D poses to search for the most likely 3D representation of a particular 2D pose [2, 9, 15].

In the deep learning context, Martinez *et al*. [22] present a fully supervised approach to predict 3D positions given 2D joint locations using a fully connected network with residual blocks. This network structure has become popular, and subsequent unsupervised approaches [3, 5, 38, 41] incorporate it within their processes. For instance, Drover *et al*. [5] propose a weakly supervised approach that lifts a 2D pose to 3D and then evaluates its 2D projection through a GAN loss. Later, Chen *et al*. [3] extended this work by adding a symmetrical pipeline of consecutive transformations (lifting, rotation, and projection) of the estimated 3D representation. This cycle of transformations exploits geometric consistency and removes the dependency on any 3D correspondences.

More recently, Wandt *et al*. [38] incorporate two fundamental elements to the model in [3] that increase the performance of the 3D lifting process: the use of normalising flow (NF) and a learned elevation angle for the 3D rotations. Previous methods have successfully used normalising flow to estimate 3D prior distributions given 3D human poses [40]. However, the method in [38] is the first to use normalising flow to infer the probability of a reconstructed 3D pose from a prior distribution of the 2D input.

Unlike these previous methods, we do not assume access to ground truth 2D poses as input. Instead, our model takes a single image and predicts the 2D pose from it, which is then lifted to 3D. Overall, it estimates both the 2D and 3D poses from the input image, removing the dependency on paired 2D pose annotations or pre-trained 2D pose predictors.

3D pose from an image

Work under this category is more related to our approach. Typically, methods for estimating 3D pose from single images break down the task into two steps. First, the 2D joints are localised, and then the 3D pose is estimated from these 2D joint positions. Early deep-learning implementations of this two-step process depend on having access to 2D ground truth joint locations and 3D data for supervising training [20, 29, 44]. Furthermore, most of those methods integrate pre-trained 2D pose estimators, e.g. a stacked hourglass network [27] and AlphaPose [6], for solving the joint localization step.

On the unsupervised side, many approaches incorporate specific assumptions for the 2D and 3D joint configuration or even add some small portion of actual 3D data to guide the training. For instance, [35] shows a unified multi-stage CNN architecture to estimate 2D and 3D joint locations from single images. This approach relies on a probabilistic 3D model of the human pose responsible for lifting the 2D representations. Other methods incorporate motion information; for example, Zhou *et al*. [45] use sequences of images and their corresponding 2D pose to guide their 3D pose estimation framework.

Like our work, Kundu *et al*. [17] propose a self-

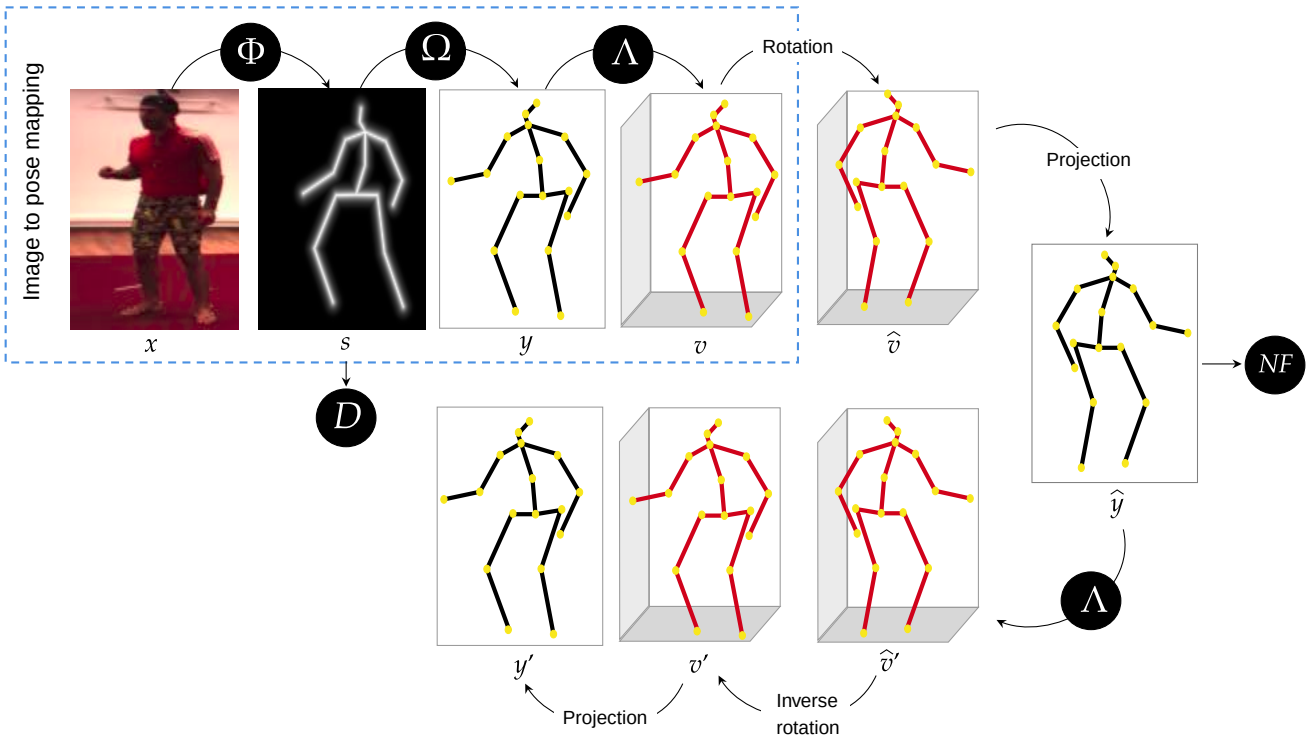

Figure 2. **Self-supervised architecture for estimating the 3D pose of a person.** Our method aims to map an image of a person x to its 3D pose v. To achieve this; we train the networks Φ, Ω, D, and Λ with help of a normalising flow network and a couple of rotations and projections. The pipeline starts with passing an input image x through Φ to obtain a skeleton image s, then Ω produces a set of 2D joint positions y given s. We then embed the mapping into a larger cyclic architecture, where the 2D pose y is lifted into 3D using Λ. The 3D pose v is then rotated, $\hat{v}$ projected into 2D, $\hat{y}$ is lifted back again into 3D through the same Λ, $\hat{v}'$ inversely rotated, and finally v' projected back to the 2D pose y' which should resemble the original 2D pose y.

supervised architecture to learn 3D poses from unlabelled images. They incorporate three assumptions: human pose articulation constraints, a part-based 2D human puppet model, and unpaired 3D poses. Other approaches explore learning without direct supervision by producing synthetic multi-views of the same skeleton [31], accessing multi-view videos [25], or relying solely on 3D kinematic constraints [18]. In contrast, our method does not depend on multi-view images nor any 3D data; it requires only an unlabelled dataset of images depicting people in typical poses and an unpaired empirical prior for the 2D pose.

3. Method

Our proposed 3D pose estimation model consists of a pipeline of three networks Φ, Ω, Λ mapping from full body images to 3D pose. This is shown in the upper-left blue dotted box in Figure 2. The pipeline consists of:

- a Convolutional Neural Network (CNN) Φ mapping from an input image x to an intermediate skeleton image s,
- a second CNN Ω mapping from s to a 2D pose representation y, and
- a fully connected network Λ lifting the 2D pose y to the required 3D pose v.

The 3D pose is represented as an articulated structure of 3D line segments corresponding to the parts of the body (e.g., torso, head, upper arm, foot).

We train the three networks together by incorporating into a larger network (Figure 2) and optimise end-to-end. This larger network is constructed to incorporate a loop of transformations of the 3D pose. The degree of geometric consistency around the loop contributes to a loss function and provides self-supervision of the training. The training starts with a dataset of images depicting people in different poses. We also assume we have a (normally unrelated) dataset of typical 2D poses from which we obtain skeleton images using a differentiable rendering function κ. These will be used in a GAN framework D to help ensure the generated skeleton images are realistic. In the following sections, we provide more details about the components of our model.

Image to 3D pose mapping

The image-to-pose mapping is the composition of networks Φ, Ω, Λ to map an image x showing a person to its 3D pose representation v. The first part of the mapping is a CNN Φ, which maps from the image x to a skeleton image $s = \Phi(x)$ showing the person as a stick figure. Our network Φ adopts a similar architecture to the autoencoder in [14] but without the decoder stage. After training, the emergent skeleton in s aligns with the person in x as expected.

Then, the network Ω maps the skeleton image s to a 2D pose representation $y = \Omega(\Phi(x))$. Informally, Ω learns to extract 2D joint positions (x_i, y_i) from the skeleton image. Finally, Λ is a neural network that lifts the 2D pose to the required pose v in 3D. In particular, $\Lambda(y)$ estimates the depth $z_i = d_i + \Delta$ for each pair of (x_i, y_i) joint positions in the input y, where Δ is a constant depth. Then, the 3D position of joint v_i in the 3D pose v is given by

$$v_i = (x_i \cdot z_i, y_i \cdot z_i, z_i) \tag{1}$$

where z_i is forced to be larger than one, to prevent ambiguity from negative depths. In line with previous works [3, 38, 41], Δ is fixed to 10.

Our lifting network Λ is based on the work of [3, 22] and extended following [38]. In this context, our extended version not only outputs the depth z_i for each joint position (x_i, y_i) in the input, it also produces a value for the elevation angle α. This angle will be used when performing the rotations of the 3D pose v. In particular, we use α to fix the elevation angle of the vertical axis to the ground-plane about which the rotation is performed.

Skeleton images and discriminator

We encourage the training network to generate realistic skeleton images with the help of an empirical prior of 2D poses. Note that these 2D poses are unpaired, i.e., they are not annotations of the training images.

The 2D poses from our empirical prior are first rendered as skeleton images using the renderer proposed by [14]. Let C be a set of connected joint pairs (i, j), e an image pixel location, and u a set of (x, y) 2D coordinates of body joint positions. The skeleton image rendering function is given by:

$$\kappa(u)_e = \exp\left(-\gamma \min_{(i,j)\in C, r\in[0,1]} ||e - ru_i - (1-r)u_j||^2\right) \tag{2}$$

Informally κ works by defining a distance field from the line segments linking joints and applies an exponential falloff to create the image.

Following [14], we use a discriminator network D that uses the prior skeleton images to encourage the predicted skeleton images to represent plausible poses. The task of D is to determine whether or not a skeleton image $s = \Phi(x)$ looks like an authentic skeleton image such as those in the prior $w = \kappa(u)$. Formally, the goal is to learn $D(s) \in [0, 1]$ to match the reference distribution of skeleton images $p(w)$ and the distribution of predicted skeleton images $q(s)$. An adversarial loss [8] compares the unpaired samples w and the predictions s:

$$\mathcal{L}_D = \mathbb{E}_w(log(D(w)) + \mathbb{E}_s(log(1 - D(s)) \tag{3}$$

Random rotations and projections

A fundamental component of our model is the lifting process which allows learning an accurate 3D pose v from the estimated 2D input y. To provide self-supervision of the lifting function and ultimately the whole end-to-end network, we emulate a second virtual view of the 3D pose v by randomly rotating it $\hat{v} = R * v$. Previous work [3] has selected a rotation matrix R by uniformly sampling azimuth and elevation angles from a fixed distribution. Recently, [38] demonstrates that learning the distribution of elevation angles leads to better results. Thus, we follow their approach and use Λ to also predict the elevation angle for the rotation matrix. The rotation around the azimuth axis R_a is sampled from a uniform distribution $[-\pi, \pi]$.

In line with [38], we also predict the dataset's normal distribution of elevation angles R_e by calculating a batch-wise mean μ_e and standard deviation σ_e. We sample from the normal distribution $\mathcal{N}(\mu_e, \sigma_e)$ to rotate the pose in the elevation direction R_e. Then, the complete rotation matrix R is given by $R = R_e^T R_a R_e$.

After rotating the 3D pose, we project $\hat{v}$ through a perspective projection. Then, the same lifting network $\Lambda(\hat{y})$ produces another 3D pose $\hat{v}'$ which is then rotated back to the original view. The final 3D pose v' is projected to 2D using the same perspective projection. This loop of transformations of the 3D pose provides a self-supervised consistency loss. In this context, we assume that if the lifting network Λ accurately estimates the depth for the 2D input y, then the 3D poses $\hat{v}$ and $\hat{v}'$ should be similar. The same principle applies to y and the final 2D projection y'. This gives the following two components of the loss function:

$$\mathcal{L}_{3d} = ||\hat{v}' - \hat{v}||^2 \tag{4}$$

$$\mathcal{L}_{2d} = ||y' - y||^2 \tag{5}$$

Additionally, the 3D poses v and v' should be similar. However, instead of comparing with an L_2 loss, we follow [38, 41] and measure the change in the difference in 3D pose between two samples from a batch at corresponding stages in the network.

$$\mathcal{L}_{def} = ||(v'^{(j)} - v'^{(k)}) - (v^{(j)} - v^{(k)})||^2 \quad (6)$$

Similar to Wandt *et al*. [38], we do not assume samples are from the same video sequence; the samples j and k may come from different sequences and subjects.

Empirical prior on 2D pose

Like Wandt *et al*. [38], we use a normalizing flow to provide a prior over 2D pose. A normalising flow transforms a simple distribution (e.g. a normal distribution) into a complex distribution in such a way that the density of a sample under this complex distribution can be easily computed.

Let $Z \in \mathbb{R}^N$ be a normal distribution and g an invertible function $g(z) = \bar{y}$ with $\bar{y} \in \mathbb{R}^N$ as a projection of the 2D human pose vector $\hat{y}$ in a PCA subspace. By a change of variables, the probability density function for $\bar{y}$ is given by

$$p_Y(\bar{y}) = p_Z(f(\bar{y}))\left|\det\left(\frac{\delta f}{\delta \bar{y}}\right)\right| \quad (7)$$

where f is the inverse of g and $\frac{\delta f}{\delta \bar{y}}$ is the Jacobian of f. Following the normalising flow implementation in [38], we represent f as a neural network [4] and optimise over a dataset of 2D poses with negative log likelihood loss:

$$\mathcal{L}_{NF} = -\log(p_Y(\bar{y})) \quad (8)$$

Additional losses

We compute a loss from the mapping from skeleton images to 2D pose $y = \Omega(s)$. We use the same loss as [14], but without pretraining Ω, i.e., we learn this mapping simultaneously with all the other networks. $\mathcal{L}_\Omega$ is given by

$$\mathcal{L}_\Omega = ||(\Omega(\kappa(u)) - u)||^2 + \lambda||(\kappa(y) - s)||^2 \quad (9)$$

where u represents a 2D pose from the unpaired prior, s is the predicted skeleton image, and λ is a balancing coefficient set to 0.1. The function κ is the skeleton image renderer defined in Equation 2.

Based on the proven effectiveness of incorporating relative bone lengths into pose estimation methods [21, 28, 38], we add this to impose a soft constraint when estimating the 3D pose. Following the formulation in [38], we calculate the relative bone lengths b_n for the n-th bone divided by the mean of all bones of a given pose v. We use a precalculated relative bone length $\bar{b}_n$ as the mean of a Gaussian prior. Then, the negative log-likelihood of the bone lengths defines a loss function $\mathcal{L}_{bl}$,

$$\mathcal{L}_{bl} = -\log(\prod_{n=1}^{N} \mathcal{N}(b_n|\bar{b}_n, \sigma_b)) \quad (10)$$

where N is the number of bones defined by the connectivity between joints. Note that this is a soft constraint that allows variation in the relative bone lengths between individuals.

3.1. Training

We train Φ, Ω, D, and Λ from scratch. Only the normalising flow is independently pre-trained, as indicated in [38]. The complete loss function for training our model has seven components expressed as $\mathcal{L}_D$ (Equation 3), $\mathcal{L}_\Omega$ (Equation 9), $\mathcal{L}_{2d}$ (Equation 5), $\mathcal{L}_{3d}$ (Equation 4), $\mathcal{L}_{def}$ (Equation 6), $\mathcal{L}_{NF}$ (Equation 8), and $\mathcal{L}_{bl}$ (Equation 10). For convenience in ablation studies, we group three of these loss terms and represent them as $\mathcal{L}_{base}$

$$\mathcal{L}_{base} = \mathcal{L}_{2d} + \mathcal{L}_{3d} + \mathcal{L}_{def} \quad (11)$$

Thus, the final composite loss function is defined as:

$$\mathcal{L} = \mathcal{L}_D + \mathcal{L}_\Omega + \mathcal{L}_{base} + \mathcal{L}_{NF} + \mathcal{L}_{bl} \quad (12)$$

During testing we only keep the pipeline composed of the trained Φ, Ω, and Λ networks illustrated in the upper-left box in Figure 2. Please see the supplementary section for a more detailed description of the networks and training.

4. Experiments

4.1. Datasets

Human3.6M: Human3.6M [12] is a widely used large-scale pose dataset consisting of videos of 11 subjects doing 17 activities against a static background. There are 3.6M frames depicting the human body and corresponding 3D body poses. In line with the standard protocol II on Human3.6M [12], we use frames from subjects S1, S5, S6, S7, and S8 for training. For testing, we use frames from subjects S9 and S11. We pre-processed the video data by cropping the human body on each frame and removing the background, using the bounding boxes and segmentation masks provided in the dataset.

MPI-INF-3DHP: MPI-INF-3DHP [24] is a human pose dataset containing 3D annotations. Unlike Human3.6M, this dataset incorporates studio and outdoor recordings. The dataset comprises eight subjects with two video sequences for each, doing different activities, e.g. walking, sitting, exercising, and reaching. We train our model with the images from the train split and evaluate with the given test set.

HandDB: HandDB [33] is a dataset of images showing human hands under different scenarios. For our experiments, we use part of the subset of hands generated from synthetic data, which contains 2D annotations for 21 key points: four for each of the five fingers and one for the wrist. For training and testing, we select two sequences (synth2 and synth3) of images from the four included in this subset and split them 80/20, respectively.

4.2. Metrics

Following previous methods [3, 5, 22, 25, 31], we use the standard *Protocol II* to evaluate the Human3.6M dataset [12]. *Protocol II* performs a rigid alignment between the predicted 3D pose and the 3D ground truth via the Procrustes method [7]. Then, it calculates the Mean Per Joint Position Error (MPJPE), which takes the average Euclidean distance between the ground-truth joint positions and the corresponding estimated positions across all 17 joints [12]. For simplicity, we refer to this metric as P-MPJPE (for Procrustes-MPJPE). For evaluation of the MPI-INF-3DHP dataset, we show the Percentage of Correct Keypoints (PCK), which represents the percentage of estimated joint positions within a fixed distance of $150mm$ from their respective ground truth. We also report the corresponding area under the curve (AUC).

4.3. Results

Using our trained model with the Human3.6M dataset, we predict 3D pose (consisting of 17 joint positions) for every frame in all videos of subjects S9 and S11 ($\sim$ 584k frames) and calculate the average P-MPJPE. Table 1 compares our method with the state-of-the-art 3D pose estimation methods.

Assumptions	Method	P-MPJPE(↓)
3D pose from 2D landmarks		
3D poses	Martinez *et al.* [22]	52.1
2D poses	Chen *et al.* [3]	68.0
2D poses	Drover *et al.* [5]	64.6
2D poses	Yu *et al.* [41]	52.3
2D poses	Wandt *et al.* [38]	36.7
3D pose from images		
3D Poses	Chen *et al.* [2]	114.2
3D Poses	Mitra *et al.* [25]	72.5
MV + P3D	Rhodin *et al.* [32]	128.6
MV + P3D	Rhodin *et al.* [31]	98.2
MV + 2D	Wandt *et al.* [39]	53.0
Unpaired 3D	Kundu *et al.* [17]	99.2
3D Struct.	Kundu *et al.* [18]	89.4
Unpaired 2D	Ours	96.7

Table 1. **P-MPJPE (in mm's) for all activities in Human3.6M.** MV = Multi view, Unpaired 3D = Unpaired 3D poses, 3D Struct. = 3D body structure constraints, Unpaired 2D = Unpaired 2D poses, P3D = Partial 3D poses (i.e, one portion of the 3D pose annotations available).

We include supervised [2, 25], semi-supervised [31, 32, 39], and self-supervised [17, 18] approaches that estimate the 3D pose from images. For a more comprehensive comparison, we also consider supervised [22] and unsupervised [3, 5, 34, 38, 41] methods that estimate 3D pose from 2D landmarks. The performance of our method exceeds that of some methods that rely on 3D supervision [2], multi-view images [31, 32] or priors on 3D data [17].

To demonstrate our model's generalisation performance, we evaluate using the MPI-INF-3DHP test dataset under different settings. First, we trained the model using Human3.6M. Second, we train our model with images from the MPI-INF-3DHP training set. Finally, we train with both images from the MPI-INF-3DHP training data and the training set of Human3.6M. For the first experiment, the set of 2D poses used for the normalising flow prior on 2D pose and the derived empirical prior on skeletons comes from Human3.6M, and for the rest the empirical prior comes from 2D poses on MPI-INF-3DHP. Table 2 presents the PCK and AUC scores for the different evaluation conditions.

	Method	PCK(↑)	AUC(↑)
Unpaired 3D	Kundu *et al.* [17]	83.2	58.7
3D Struct.	Kundu *et al.* [18]	79.2	43.4
Unpaired 2D	Ours^{-}	58.7	24.3
Unpaired 2D	Ours*	69.6	32.8
Unpaired 2D	Ours^{+}	75.3	40.0

Table 2. **Evaluation results on MPI-INF-3DHP dataset.** First column shows the main assumption for each method, where Unpaired 3D = Unpaired 3D poses, 3D Struct. = 3D body structure constraints, and Unpaired 2D = Unpaired 2D poses. Ours^{-} indicates the model trained with Human3.6M and tested with MPI-INF-3DHP. Ours* represents the model trained with images from MPI-INF-3DHP. Ours^{+} indicates that the MPI-INF-3DHP train set has been extended with images from Human3.6M.

Figure 3 shows qualitative results on images from Human3.6M and MPI-INF-3DHP datasets. It includes a random selection of predicted 3D poses and their corresponding input image.

To demonstrate the adaptability of the method, we applied it to estimate hand pose using part of the synthetic subset from [33]. This required a different structure for the target 3D pose. For training and testing the model, we select sequences of images showing hands under similar conditions (synth2 and synth3). We augment the training set offline by making two rotated versions of each image ($45°$ and $90°$). We use half of the 2D annotations provided with the dataset to build the prior on 2D hand poses. The training set does not include the images corresponding to those annotations. With the trained model, we estimate 3D hand poses consisting of 21 key points (representing hand joint positions). Since the synthetic subset of HandDB does not contain 3D annotations (just 2D), we only show qualitative results in Figure 4.

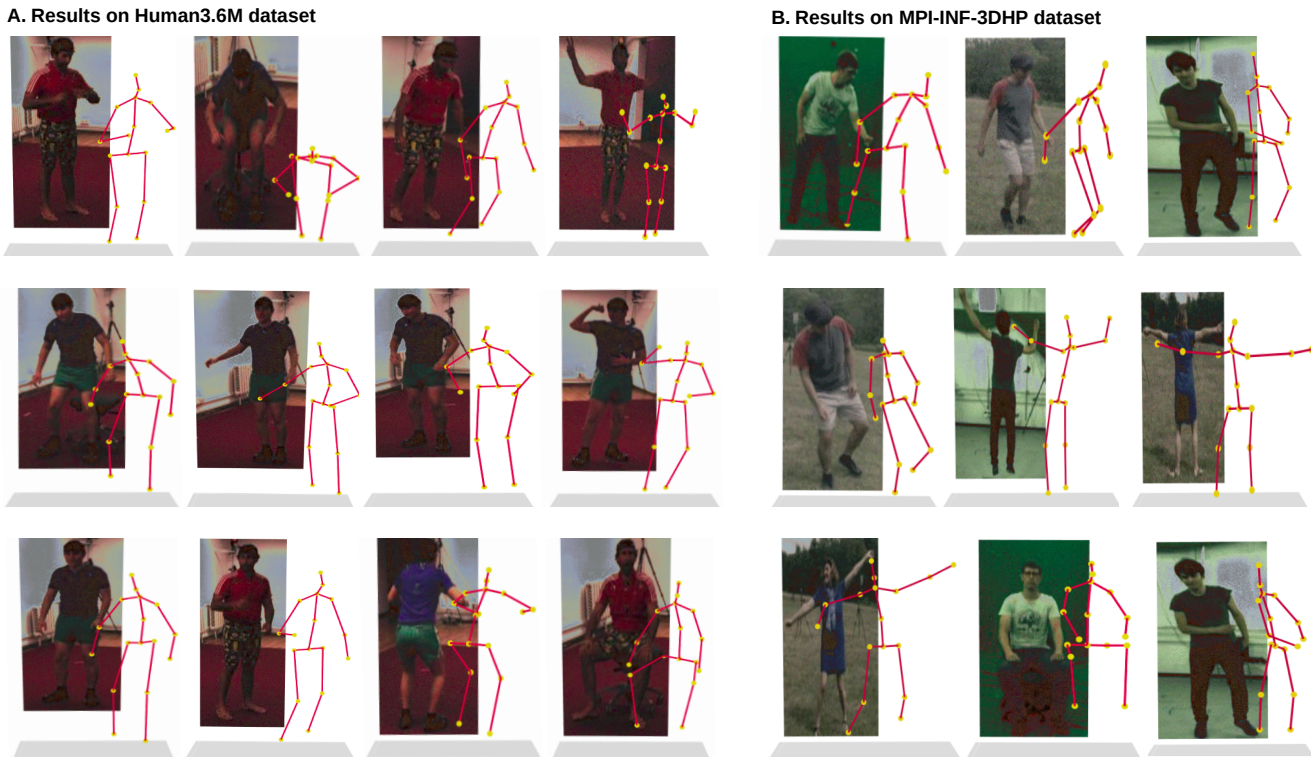

Figure 3. **Qualitative results on images from Human3.6M and MPI-INF-3DHP datasets.** Each figure contains the input image and its corresponding estimated 3D pose by our model. An extended version of qualitative results for both datasets is included in the supplemental document.

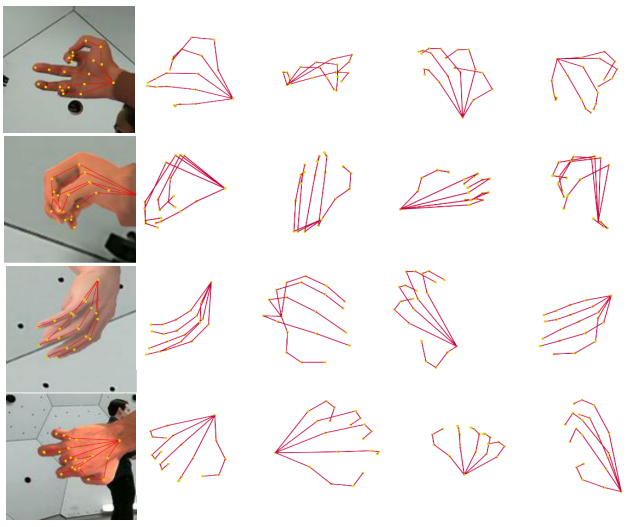

Figure 4. **Qualitative results on HandDB dataset.** First column shows the input image overlay with the 2D ground truth annotations from the dataset. Next columns display novel views of the 3D hand predictions.

4.4. Ablation study

To evaluate the effectiveness of the design of our loss function, we progressively removed components from the complete loss expressed by Equation 12. First, we evaluate the model guided by the discriminator loss $\mathcal{L}_D$ and $\mathcal{L}_\Omega$. As expected, even when the 2D predictions are mostly accurate, the overall performance decreases since there is nothing else to regulate the 3D predictions and these are more likely to be deformed. We keep the $\mathcal{L}_D$, $\mathcal{L}_\Omega$, and $\mathcal{L}_{base}$ losses for the second experiment, i.e., removing $\mathcal{L}_{NF}$ and $\mathcal{L}_{bl}$. Although the model can produce plausible 3D poses, the performance is still inferior.

Finally we assess the performance when removing only the loss term constraining the bone lengths $\mathcal{L}_{bl}$ from the original loss formulation (Equation 12). Adding the combination of the loss terms for the normalising flow prior on 2D pose $\mathcal{L}_{NF}$ and relative bone length $\mathcal{L}_{bl}$ has proven to be useful, increasing the performance of the model by 20.8% with respect to the loss function that does not contain those terms. Table 3 shows the results for each of the modifications to the loss function.

Configuration	P-MPJPE(↓)
$\mathcal{L}_D + \mathcal{L}_\Omega$	139.3
$\mathcal{L}_D + \mathcal{L}_\Omega + \mathcal{L}_{base}$	122.1
$\mathcal{L}_D + \mathcal{L}_\Omega + \mathcal{L}_{base} + \mathcal{L}_{NF}$	112.3
$\mathcal{L}_D + \mathcal{L}_\Omega + \mathcal{L}_{base} + \mathcal{L}_{NF} + \mathcal{L}_{bl}$	**96.7**

Table 3. **Ablation studies.** Experiments with different loss terms using Human3.6M dataset for training and testing.

5. Discussion

Our method outperforms self-supervised state-of-the-art approaches that estimate 3D pose from images and assume unpaired 3D data for supervision [17]. Also, it performs better than recent methods that rely on 3D supervision [2] or multi-view images [31,32]. Moreover, its performance is similar to one method that assume 3D kinematic constraints [18]. We achieve superior performance to Kundu *et al*. [18] in 20% of the activities (*Discuss, Pose,* and *Wait*), and close scores for the rest.

Most failure cases of our model on the Human3.6M dataset appear for snapshots from activities such as *Sitting* and *Sitting Down*. We assume this occurs because of the self-occlusions and perspective ambiguity in these activities. However, according to the examples shown in Figure 5, the model can still produce plausible 3D poses for most cases, even if they do not exactly match their respective 3D ground truth. The high P-MPJPE comes from mismatches between the 'joints' representing the body's extremities, e.g. hands and feet.

With the experiments using the MPI-INF-3DHP dataset, we demonstrate the cross-dataset generalisation of our method. We also show that it works well even if we train the NF network with a different dataset (Human3.6M). Moreover, when training with images from Human3.6M and MPI-INF-3DHP, the experimental results suggest that increasing the diversity of images in the training set could help to improve the overall performance. In principle, extending the dataset of images is relatively straightforward since 3D annotations are not require.

6. Conclusion

We demonstrate how to estimate 3D human pose with a training architecture requiring only images depicting people in different poses and an unpaired set of typical 2D poses. We demonstrate qualitatively that our approach holds the potential for rapidly learning about the pose of articulated structures other than the human body without the need to collect ground-truth 3D pose data, e.g. human hands.

Overall, using human datasets the qualitative and quantitative results suggest that our method is comparable to other self-supervised state-of-the-art approaches that estimate 3D pose from images despite requiring a less onerous dataset for training. Furthermore, it performs better than state-of-the-art methods that rely on multi-view images or 3D pose annotations for supervision. Prior work has demonstrated the value of using temporal information from image sequences and domain adaptation networks. Incorporating these into our approach would be a promising direction for future work. Finally, the way is open to apply the method to much larger datasets of unlabelled images to see whether performance continues to improve, and to apply the method to other articulated structures (e.g., mice, dogs and other animals), exploiting the relatively light requirement for self-supervision in the form of an empirical prior on 2D poses.

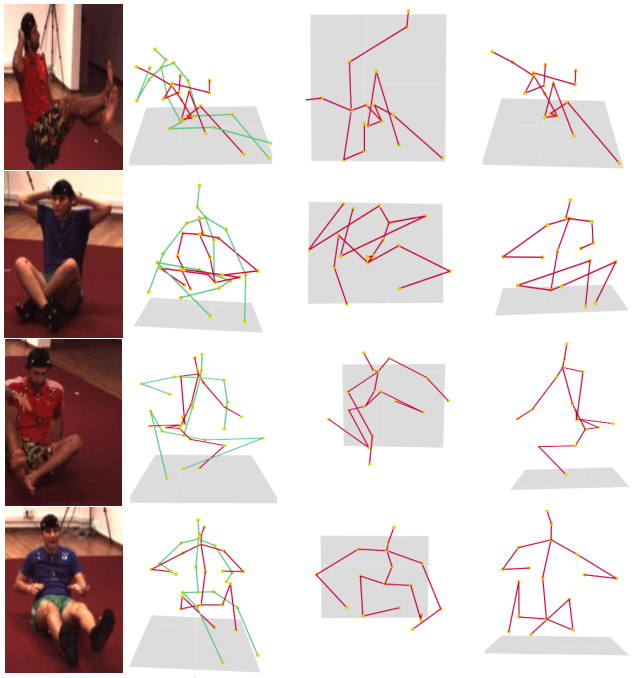

Figure 5. **Failure cases on Human3.6M.** 3D predictions with a P-MPJPE greater than 200mm. The first column shows the input images. The second column displays the predicted 3D pose (coloured in red) aligned with its respective ground truth (coloured in green). Following columns show different views of the predicted 3D pose.

References

[1] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):44–58, 2005. 1

[2] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7035–7043, 2017. 1, 2, 6, 8

[3] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Pro-*

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5714–5724, 2019. 1, 2, 4, 6

[4] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 5

[5] Dylan Drover, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2, 6

[6] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *Proceedings of the IEEE international conference on computer vision*, pages 2334–2343, 2017. 2

[7] Colin Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(2):285–321, 1991. 6

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 4

[9] Ankur Gupta, Julieta Martinez, James J Little, and Robert J Woodham. 3d pose from motion for cross-view action recognition via non-linear circulant temporal encoding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2601–2608, 2014. 1, 2

[10] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1(1):5–20, 1983. 1

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

[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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 5, 6

[13] Umar Iqbal, Pavlo Molchanov, and Jan Kautz. Weakly-supervised 3d human pose learning via multi-view images in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5243–5252, 2020. 1

[14] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Self-supervised learning of interpretable keypoints from unlabelled videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8787–8797, 2020. 4, 5

[15] Hao Jiang. 3d human pose reconstruction using millions of exemplars. In *2010 20th International Conference on Pattern Recognition*, pages 1674–1677. IEEE, 2010. 1, 2

[16] Wonhui Kim, Manikandasriram Srinivasan Ramanagopal, Charles Barto, Ming-Yuan Yu, Karl Rosaen, Nick Goumas, Ram Vasudevan, and Matthew Johnson-Roberson. Pedx: Benchmark dataset for metric 3-d pose estimation of pedestrians in complex urban intersections. *IEEE Robotics and Automation Letters*, 4(2):1940–1947, 2019. 1

[17] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6152–6162, 2020. 2, 6, 8

[18] Jogendra Nath Kundu, Siddharth Seth, MV Rahul, Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11312–11319, 2020. 1, 3, 6, 8

[19] Hsi-Jian Lee and Zen Chen. Determination of 3d human body postures from a single view. *Computer Vision, Graphics, and Image Processing*, 30(2):148–168, 1985. 2

[20] 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. 2

[21] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2192–2201, 2019. 5

[22] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2640–2649, 2017. 2, 4, 6

[23] 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 (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2

[24] 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 *2017 international conference on 3D vision (3DV)*, pages 506–516. IEEE, 2017. 5

[25] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6907–6916, 2020. 3, 6

[26] Ramakant Nevatia and Thomas O Binford. Description and recognition of curved objects. *Artificial intelligence*, 8(1):77–98, 1977. 1

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 1, 2

[28] Qiang Nie, Jiangliu Wang, Xin Wang, and Yunhui Liu. View-invariant human action recognition based on a 3d bio-constrained skeleton model. *IEEE Transactions on Image Processing*, 28(8):3959–3972, 2019. 5

[29] Sungheon Park, Jihye Hwang, and Nojun Kwak. 3d human pose estimation using convolutional neural networks with 2d pose information. In *European Conference on Computer Vision*, pages 156–169. Springer, 2016. 2

[30] Konstantinos Rematas, Ira Kemelmacher-Shlizerman, Brian Curless, and Steve Seitz. Soccer on your tabletop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4738–4747, 2018. 1

[31] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018. 1, 3, 6, 8

[32] Helge Rhodin, Jörg Spörri, Isinsu Katircioglu, Victor Constantin, Frédéric Meyer, Erich Müller, Mathieu Salzmann, and Pascal Fua. Learning monocular 3d human pose estimation from multi-view images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8437–8446, 2018. 6, 8

[33] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2, 5, 6

[34] Cheng Sun, Diego Thomas, and Hiroshi Kawasaki. Unsupervised 3d human pose estimation in multi-view-multi-pose video. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5959–5964. IEEE, 2021. 6

[35] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2500–2509, 2017. 2

[36] Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, and Christoph Bregler. Efficient object localization using convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 648–656, 2015. 1

[37] Alexander Toshev and Christian Szegedy. Deeppose: Human pose estimation via deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1653–1660, 2014. 1

[38] Bastian Wandt, James J Little, and Helge Rhodin. Elepose: Unsupervised 3d human pose estimation by predicting camera elevation and learning normalizing flows on 2d poses. *arXiv preprint arXiv:2112.07088*, 2021. 2, 4, 5, 6

[39] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13294–13304, 2021. 6

[40] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11199–11208, 2021. 2

[41] Zhenbo Yu, Bingbing Ni, Jingwei Xu, Junjie Wang, Chenglong Zhao, and Wenjun Zhang. Towards alleviating the modeling ambiguity of unsupervised monocular 3d human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8651–8660, 2021. 1, 2, 4, 6

[42] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012. 1

[43] Liang Zheng, Yujia Huang, Huchuan Lu, and Yi Yang. Pose-invariant embedding for deep person re-identification. *IEEE Transactions on Image Processing*, 28(9):4500–4509, 2019. 1

[44] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 398–407, 2017. 2

[45] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4966–4975, 2016. 2