ElePose: Unsupervised 3D Human Pose Estimation by Predicting Camera Elevation and Learning Normalizing Flows on 2D Poses

Bastian Wandt, James J. Little, and Helge Rhodin
University of British Columbia
{wandt,little,rhodin}@cs.ubc.ca

Abstract

Human pose estimation from single images is a challenging problem that is typically solved by supervised learning. Unfortunately, labeled training data does not yet exist for many human activities since 3D annotation requires dedicated motion capture systems. Therefore, we propose an unsupervised approach that learns to predict a 3D human pose from a single image while only being trained with 2D pose data, which can be crowd-sourced and is already widely available. To this end, we estimate the 3D pose that is most likely over random projections, with the likelihood estimated using normalizing flows on 2D poses. While previous work requires strong priors on camera rotations in the training data set, we learn the distribution of camera angles which significantly improves the performance. Another part of our contribution is to stabilize training with normalizing flows on high-dimensional 3D pose data by first projecting the 2D poses to a linear subspace. We outperform the state-of-the-art unsupervised human pose estimation methods on the benchmark datasets Human3.6M and MPI-INF-3DHP in many metrics.

1. Introduction

Human pose estimation from single images is an ongoing research topic with many applications in medicine, sports, and human-computer interaction. Tremendous improvements have been achieved in recent years via machine learning. However, many recent approaches rely on a large amount of data used to train a 3D pose estimator in a supervised fashion. Unfortunately, such training data is hard to record and rarely available for specialized domains. For this reason, recent work focuses on reducing the amount of labeled data by using weak supervision in the form of unpaired 2D-3D examples, sparse supervision with a small amount of labeled 3D data, or multi-view setups during training. In contrast, we propose a method that is trained only from 2D data, which is easy to annotate by clicking visible keypoints in readily available images and thereby alleviates the 3D labelling and multi-view capture steps required by weakly- and fully-supervised approaches.

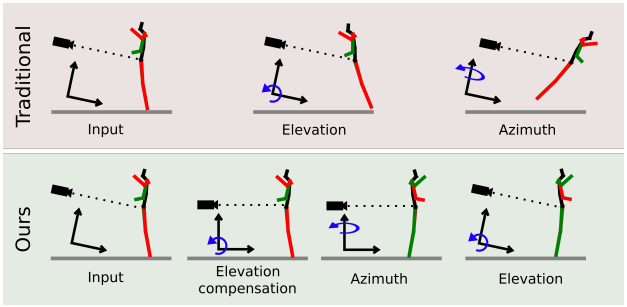

Figure 1. The elevation sampling for the creation of virtual views performed by traditional methods and by our approach are shown from left to right. The commonly used prior distribution of the randomly sampled elevation angle leads to errors if it does not exactly match the distribution in the training dataset. ElePose solves this problem by learning this distribution and compensating for it before applying other transformations which significantly increases the performance of our method.

Given observations of 2D human joints in monocular imagery, we train a neural network to recover the depth —the *missing* third coordinate of the 3D human pose. With the same goals, Chen et al. [4] and Yu et al. [63] train a 3D pose estimator in an adversarial setting [11]. Their generator predicts a 3D pose that is randomly rotated and projected to a virtual camera which is fed into an adversarial network over 'fake' projected 3D poses and the 'real' 2D pose distribution. The idea is that, for a correct 3D pose prediction, the rotated and projected 2D pose should also come from the distribution of 2D training poses. However, the predicted 3D pose is rotated randomly over a fixed prior distribution defined relative to the camera coordinate system. It is a reasonable assumption when the camera is close to parallel to the ground plane. However, even for small elevation angles and even when modeling variation by sampling from a predefined Gaussian distribution, this leads to

random projections that cannot be found in the training data as shown in Fig. 1.

We build upon this concept and improve the handling of varying camera angles. Our core contribution is to train a network that predicts the elevation for every 2D input. After correcting for the predicted elevation, 3D reconstructions are upright such that rotating around the y-coordinate corresponds to rotating around the up-direction and uniformly sampling from all possible azimuth angles is meaningful as human poses are generally symmetric around the direction of gravity and the ground normal. While camera angles have been estimated in supervised and weakly supervised settings [13, 14, 19, 56, 57] we do it for the monocular case and without supervision.

The approach of projecting to random virtual cameras requires to know the distributions of camera poses. Tailored to this, we propose a method for estimating the distribution of elevation angles from multiple point estimates, which further improves the performance of our model.

Another major change compared to previous work is that we use normalizing flows to learn a prior distribution over 2D poses, which is subsequently used to infer the most likely up-to-scale 3D pose. By contrast to GANs, which at best give a surrogate to the likelihood of an outcome with the discriminator response, our probabilistic formulation via normalizing flows naturally gives a likelihood for a predicted pose during inference time. Besides gaining a significant improvement in accuracy and robustness over existing methods, our approach is also able to provide a measure for its performance, which is very valuable information in practical applications.

We overcome several technical challenges to make training and inference tractable. First, the bijectivity of normalizing flows is a useful property, which enables them to avoid mode collapse. However, their construction restricts their input and output dimensions to be equal. For high-dimensional data, such as human poses, this leads to non-optimal convergence and an incomplete latent space. Second, the normalizing flow is still an approximation to the true pose distribution and can predict a high likelihood for poses that are outside the training distribution. Optimizing the depth estimation network to produce 3D poses with high likelihoods for their back projections causes convergence to non-optimal solutions. To avoid this, we propose to first project the 2D poses to a lower-dimensional space given by a Principal Components Analysis (PCA) on the training data. Additionally, we introduce a suitable prior for the relative bone lengths in the human body to predict anthropometrically valid 3D poses.

Ethics and general impact. Building such an unsupervised approach for motion capture promises to be more inclusive to people and activities that are not well represented in current motion capture datasets. Source code: https://github.com/bastianwandt/ElePose.

Pose estimators could be abused for unwanted surveillance and our method could be used for motion pattern analysis. However, we believe this risk is low since it does not reconstruct any visual features.

2. Related Work

In this section we discuss recent 3D human pose estimation approaches, structured by the different types of supervision, and put our approach in context.

Full Supervision. Supervised approaches rely on large datasets that contain millions of images with corresponding 3D pose annotations. Li et al. [25] were the first to apply CNNs to regress a 3D pose from image input directly. They later improved their work [27] by a structured learning framework. Others followed this image-to-3D approach [9, 17, 26, 29, 33, 37, 39, 43, 48, 50–53, 62, 65]. Typically, these end-to-end approaches achieve exceptional performance on similar image data. However, they struggle to generalize to very different scenes. To avoid the dependence on image data other approaches use a pretrained 2D joint detector [2, 10, 31, 32, 35, 38]. Martinez et al. [30] train a neural network on 2D poses and corresponding 3D ground truth. Due to its simplicity, it can be trained quickly for a large number of epochs leading to high accuracy, and serves as a baseline for many following approaches. While effective, the major downside of all supervised methods is that they do not generalize well to images with unseen poses.

Weak Supervision. Weakly supervised approaches require only a small set of labeled 3D poses or unpaired 2D and 3D poses. Several approaches assume unpaired 2D and 3D poses [6, 12, 22, 56, 58, 64] and leverage available motion capture data and combine it with unknown 2D data. To allow for in-the-wild pose estimation of datasets where no training data is available, a transfer learning approach is introduced by Mehta et al. [31] which was later improved in Mehta et al. [33] to achieve real-time performance. Other work first learns an embedding of multi-view data which is then used to train a 3D pose estimator with a sparse set of labeled 3D poses. Rhodin et al. [45, 46] use multi-view images and known camera positions to learn a 3D pose embedding. Others [34, 36, 44, 49] followed the same idea. Compared to completely supervised approaches, these weakly supervised methods generalize and transfer better to new domains. However, they still struggle with poses that are very different from the labeled training set.

Multi-view Supervision without 3D Data. Multi-view approaches only use information from multiple cameras

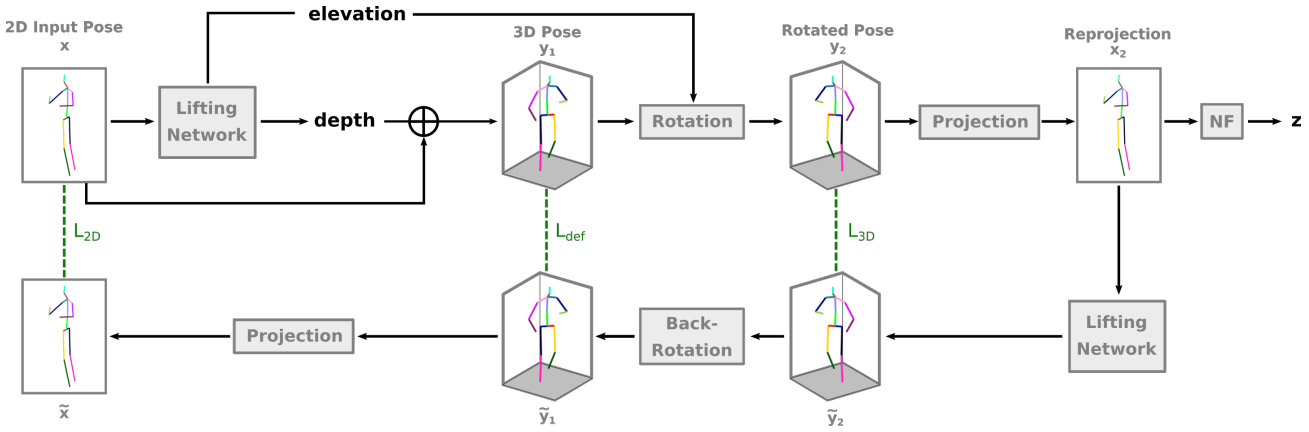

Figure 2. Overview of our approach. Given a normalized 2D input pose a lifting network predicts a depth for each joint coordinate which gives a 3D pose. Additionally it predicts the camera elevation in a parallel path. This 3D pose is randomly rotated and projected to 2D. The pretrained normalizing flow computes the negative log-likelihood which is used as a loss to train the lifting network.

without requiring any 3D data. Rochette et al. [47] achieve similar performance as a comparable fully supervised approach. However, they use a large number of cameras from different viewing angles, which limits the practical applicability. Kocabas et al. [20] apply epipolar geometry to 2D poses from multiple views to compute a pseudo ground truth which is then used to train the 3D lifting network. Iqbal et al. [16] train an end-to-end network that refines the pre-trained 2D pose estimator during the self-supervised training. Likewise, Wandt et al. [57] reconstruct 3D poses in a canonical pose space that is consistent over all views. While these multi-view approaches are a promising direction towards motion capture in the wild, they still require multiple temporally synchronized cameras for training.

Unsupervised. This section covers work that does not use any 3D data or additional views. Our work also falls into this category. Drover et al. [8] propose an unsupervised learning approach to monocular human pose estimation. They randomly project an estimated 3D pose back to 2D. This 2D projection is then evaluated by a discriminator following adversarial training approaches. However, they create an artificial 2D dataset from known ground-truth 3D poses. Chen et al. [4] extend [8] with a cycle consistency loss that is computed by lifting the randomly projected 2D pose to 3D and inversing the previously defined random projection. In contrast to Drover et al. [8] they only use 2D data given by the dataset. Yu et al. [63] build upon [4] and introduce a learnable scaling factor for the input 2D poses. None of them estimates the camera orientation and its distribution and we are also the first to apply normalizing flows to this setting. Once trained, unsupervised methods do not generalize better than supervised ones, but they can sidestep this problem by training on examples taken from the target domain as no 3D labelling is required.

Normalizing Flow. Informally, a normalizing flow is a tool to efficiently map distributions back and forth between two spaces. It applies to probability density estimation, which we use for the likelihood estimation of poses.

Let $\mathcal{Z} \in \mathbb{R}^N$ be a known distribution (in our case a normal distribution) and g be an invertible function $g(\mathbf{z}) = \mathbf{x}$, with $\mathbf{x} \in \mathbb{R}^N$ as a vector representing the joints of a human pose[1]. With the change of variables formula the probability density function of $\mathbf{x}$ is computed as

$$p_{\mathcal{X}}(\mathbf{x}) = p_{\mathcal{Z}}(f(\mathbf{x})) \left| \det \left(\frac{\partial f}{\partial \mathbf{x}} \right) \right|, \tag{1}$$

where f is the inverse of g and $\frac{\partial f}{\partial \mathbf{x}}$ is the Jacobian of f. That means given an invertible function f the density of a 2D pose $\mathbf{x}$ can be calculated by the product of the density of its projection $f(\mathbf{x})$ with the respective Jacobian determinant. In our case f is the trainable neural network proposed in [7]. Details on the construction and training are given in the supplemental document. Normalizing flows have been used to learn prior distributions of 3D human poses [1,21,61,64] or to model ambiguities during the lifting step [59]. However, they aim to build a probabilistic 3D model of a skeleton and therefore require 3D training data. To the best of our knowledge, our approach is the first that uses normalizing flows to learn the prior distribution of 2D input data to infer the probability of a reconstructed 3D pose.

3. Formulation

Our goal is to train a neural network that given a root-centered 2D pose $\mathbf{x} \in \mathbb{R}^{2J}$ recovers the 3D pose $\mathbf{y} \in \mathbb{R}^{3J}$ with J 3D joint positions. In our unsupervised setting, only a dataset of 2D poses is available for training. The general

[1]This could be either the 2D pose vector $\mathbf{x}$ or its image in the PCA subspace.

concept follows the assumptions of previous work in this domain, i.e., we assume that a 3D pose is plausible when viewing its 2D projections from multiple views. The difficulty lies in finding a plausibility measure for these 2D projections without multi-view data. We propose to learn this measure via normalizing flows. As a second major contribution we learn the camera angle distribution instead of predefining a dataset-dependent prior. This not only makes our approach more flexible but also overcomes the problem of wrongly rotated poses resulting in significant improvement in performance. All parts of our pipeline are visualized in Fig. 2 and explained in the following sections.

Lifting and Camera Model. Given 2D joint locations $\mathbf{x} \in \mathbb{R}^{2\times J}$, we introduce a lifting network that predicts for every joint j the depth $\mathbf{w}_j = \mathbf{d}_j + D$ as the offset $\mathbf{d}_j$ to a constant depth D. The full 3D pose is reconstructed based on the perspective unprojection

$$\mathbf{y}_j = [\mathbf{u}_j \mathbf{w}_j, \mathbf{v}_j \mathbf{w}_j, \mathbf{w}_j], \tag{2}$$

where $\mathbf{u}_j$ and $\mathbf{v}_j$ are the horizontal and vertical joint positions in the image. It inverts the perspective projection operation

$$P(\mathbf{y}_j) = P([\mathbf{y}_j^{(\mathrm{x})}, \mathbf{y}_j^{(\mathrm{y})}, \mathbf{y}_j^{(\mathrm{z})}]) = [\mathbf{y}_j^{(\mathrm{x})}/\mathbf{y}_j^{(\mathrm{z})}, \mathbf{y}_j^{(\mathrm{y})}/\mathbf{y}_j^{(\mathrm{z})}], \tag{3}$$

with $[\mathbf{y}_j^{(\mathrm{x})}, \mathbf{y}_j^{(\mathrm{y})}, \mathbf{y}_j^{(\mathrm{z})}]$ as the 3D position of joint j. To prevent ambiguous reconstructions with negative depth, $\mathbf{w}_j$ is clipped to be larger than one. Following previous work [4,63] the depth D is fixed to $D = 10$, as perspective effects change little with depth, and each 2D pose $\mathbf{y}$ is normalized by centering it at the root joint and dividing it by the mean length of the vector from the root joint to the head joint.

Reprojection to Virtual Cameras. We motivate our approach from multi-view camera setups, where the depth can be supervised by reprojection to the other views. Since no multi-view data is available in an unsupervised setting we assume a *virtual* second view. It requires rotating 3D poses centered at their root joint with $\mathbf{y}_2 = \mathbf{R}[\mathbf{y}_1]^{3\times J}$, where $\mathbf{R} \in \mathbb{R}^{3\times 3}$ is the rotation matrix from the original camera to a virtual camera and $[\mathbf{y}_1]^{3\times J}$ the pose vector $\mathbf{y}_1$ in the original camera coordinate system reshaped to a matrix with one of the J 3D joint positions in each column. Typically, the rotation $\mathbf{R}$ is randomly sampled from a predefined distribution $\mathcal{R}$ [4, 63]. However, in general, it is unknown and different for every dataset. One of our core contributions is to learn this distribution instead of predefining it which we discuss in the following. Using the same perspective camera model as for the lifting in Eq. 3 the 2D pose $\mathbf{x}_2 = P(\mathbf{y}_2)$ is computed by moving the predicted 3D pose with the predefined translation D and dividing each joint by its depth.

Reprojection Likelihood In a multi-view setting the reprojection likelihood is typically a Gaussian with standard deviation σ_r, centered at the 2D projection $\mathbf{x}_2$ of the 3D pose $\mathbf{y}_2$ leading to a least squares loss when inferred using maximum likelihood or MAP. Since multi-camera information is not available in an unsupervised setting there exists no corresponding 2D pose that can be matched to $\mathbf{x}_2$. While previous work [4, 63] tries to learn the distribution of plausible 2D poses with an adversarial approach we leverage normalizing flows for learning the probability density function of the 2D poses in the training dataset. We define the reprojection likelihood by computing the likelihood of the latent variable $\mathbf{z}$ in the latent space of a normalizing flow using Eq. 1. In contrast to [4, 63], this enables us to compute a likelihood for each reconstructed 3D pose which is very valuable information for downstream tasks.

In practice, we minimize the negative log-likelihood of Eq. 1 which gives the normalizing flow loss

$$\mathcal{L}_{NF} = -\log(p_{\mathcal{X}}(\mathbf{x})). \tag{4}$$

Stabilized Normalizing Flows. We found that directly training the normalizing flow on 2D poses leads to non-optimal convergence during training of the lifting network. We hypothesize that it is due to the high dimensionality of the input data which leads to a sparse latent space of the normalizing flow. That means the latent space contains poses that are not in the original distribution of 2D poses, although the normalizing flow assigns a high likelihood to them. To mitigate this, instead of directly estimating the likelihood of a 2D pose we propose to first project the 2D pose to a low dimensional subspace. Our subspace is determined by principal components analysis. The projection to the subspace eliminates redundancies and noise from the data and, therefore, leads to a more stable training of the normalizing flow and subsequently the lifting network.

Camera Distribution and Elevation. The predicted 3D pose $\mathbf{y}$ is rotated to a virtual view by randomly sampling $\mathbf{R} \sim \mathcal{R}$. To achieve a reasonable 2D projection of the rotated 3D pose the distribution $\mathcal{R}$ needs to be defined such that it matches the distribution of rotations present in the training data set. In general, $\mathcal{R}$ is unknown in an unsupervised setting. However, there are reasonable priors for camera setups based on natural human behaviour while recording another human: 1) cameras are held horizontally, 2) since gravity defines a clear up-direction, cameras (or observed subjects) are mostly rotated around the azimuth axis, and 3) similar activities are recorded with similar but slightly varying elevation angles. In terms of $\mathcal{R}$ these three points mean that 1) there is negligible rotation around the optical axis, 2) a uniform prior over 360° rotation around the azimuth axis is plausible, and 3) an

unknown but restricted rotation around the elevation axis. While the former two assumptions can be straightforwardly modeled the latter is commonly approximated by sampling the elevation angle from a uniform distribution in the interval $[-\pi/9, \pi/9]$ [4, 63]. Unfortunately, this can lead to situations where a reconstructed person is strongly tilted towards the camera as visualized in Fig. 1. This in turn results in backprojections that cannot be observed in the training set. As a major contribution in this paper we propose to learn the distribution of the elevation angle $\mathcal{R}_e$. Since each 2D pose can have a unique elevation angle the lifting network is extended by a branch that predicts the elevation angle. The resulting rotation matrix $\mathbf{R}_e$ is used to rotate the predicted 3D pose $\mathbf{y}$ to the direction of gravity by $\mathbf{R}_e^T[\mathbf{y}]^{3\times J}$. This step alone already improves the predictions since it compensates for the formerly ignored elevation and therefore the azimuth rotation is correctly applied around the azimuth axis.

To further improve the results we additionally use the elevation predictions to predict the normal distribution of elevation angles in the dataset by calculating the mean μ_e and standard deviation σ_e over all elevation angles in a batch such that

$$p(\mathcal{R}_e) = \mathcal{N}(\mu_e, \sigma_e). \tag{5}$$

The rotation around the azimuth axis $\mathbf{R}_a$ is randomly sampled from a uniform distribution in the interval $[-\pi, \pi]$. To rotate the pose back in elevation direction a rotation $\tilde{\mathbf{R}}_e$ for each sample in the batch is randomly sampled from the normal distribution $\mathcal{N}(\mu_e, \sigma_e)$. To allow for backpropagation through the sampling step we use the same reparametrization as for variational autoencoders, *i.e.*

$$\tilde{\mathbf{R}}_e \sim \mu_e + \sigma_e \mathcal{N}(0, 1). \tag{6}$$

The full rotation $\mathbf{R}$ can now be written as

$$\mathbf{R} = \mathbf{R}_e^T \mathbf{R}_a \tilde{\mathbf{R}}_e. \tag{7}$$

Our experiments show that our novel elevation angle estimation significantly improves results by approximately 15% in PA-MPJPE and more than 22% in MPJPE.

Skeleton likelihood. Human poses have several anthropometric properties defined by the kinematic chain of bones. Most of these properties, such as bone lengths and joint angle limits, are unknown in an unsupervised setting. However, *relative* bone lengths are nearly constant across people [42]. For this reason, we calculate the relative bone lengths b_k for the k-th bone divided by the mean length of all bones of a single pose. We use a Gaussian prior with the mean at the pre-calculated relative bone length $\bar{b}_k$. The density for the bone lengths prior is given by

$$p(b_1, b_2, \ldots, b_K | \bar{b}_1, \bar{b}_2, \ldots, \bar{b}_K) = \prod_{k=1}^{K} \mathcal{N}(b_k | \bar{b}_k, \sigma_b), \tag{8}$$

where K is the number of bones. This forms a prior in terms of 3D pose $\mathbf{y}$ and a likelihood, $p(\mathbf{x}_1, \mathbf{d})$, of $\mathbf{x}_1$ given a depth $\mathbf{d}$ since a 3D pose is formed as a combination of observation and latent variable. Practically, we define the loss $\mathcal{L}_{bone}$ as the negative log-likelihood of Eq. 8. Note that our formulation imposes a soft constraint but does not fix bones to a predefined length.

Additional Losses. We additionally employ 3 losses similar to [63], namely the 3D lifting loss $\mathcal{L}_{3D}$, the deformation loss $\mathcal{L}_{def}$, and the 2D reprojection loss $\mathcal{L}_{2D}$. Figure 2 visualizes these three losses. Since the 3D pose $\mathbf{y}_2$ that produces the 2D pose $\mathbf{x}_2$ is known the lifting network is applied again to $\mathbf{x}_2$ to obtain the lifted pose $\tilde{\mathbf{y}}_2$. We define the traditional supervised L_2 loss

$$\mathcal{L}_{3D} = \|\tilde{\mathbf{y}}_2 - \mathbf{y}_2\|_2. \tag{9}$$

By rotating $\tilde{\mathbf{y}}_2$ back to the original view we get a 3D pose $\tilde{\mathbf{y}}_1 = \mathbf{R}^T \tilde{\mathbf{y}}_2$ that should match $\mathbf{y}_1$. Yu *et al.* [63] showed that instead of directly applying another L_2 loss on these two poses it is beneficial to consider the deformation between two poses at different time steps. Since we do not assume any temporal data we define the same loss between two samples of a batch that could come from different people and sequences. For the poses $\mathbf{y}_1$ and $\tilde{\mathbf{y}}_1$ at batch position a and b we define the time and pose independent deformation loss

$$\mathcal{L}_{def} = \|(\tilde{\mathbf{y}}_1^{(a)} - \tilde{\mathbf{y}}_1^{(b)}) - (\mathbf{y}_1^{(a)} - \mathbf{y}_1^{(b)})\|_2. \tag{10}$$

Using the same perspective projection as before $\tilde{\mathbf{y}}_1$ is projected to the 2D pose $\tilde{\mathbf{x}} = P(\tilde{\mathbf{y}}_1)$. This gives the 2D back projection loss

$$\mathcal{L}_{2D} = \|\tilde{\mathbf{x}} - \mathbf{x}\|_1. \tag{11}$$

Since the combination of these three terms has proven to be successful in [63] we summarize them as our basis loss

$$\mathcal{L}_{base} = \mathcal{L}_{3D} + \mathcal{L}_{def} + \mathcal{L}_{2D}. \tag{12}$$

Neural Network Structure. The lifting network is inspired by the MLP-based lifting network from Martinez et al. [30] and consists of 3 residual blocks, each of which contains 2 fully connected layers with 1024 neurons followed by a leaky ReLU activation function. The input is upscaled to a dimension of 1024 by a fully connected layer followed by a leaky ReLU activation function. Downscaling to the dimension of the depth is performed by another fully connected layer without activation. The elevation angle is predicted in a path parallel to the 3 residual blocks of the depth estimation network that has identical architecture. The normalizing flow consists of 8 coupling blocks. Each sub-network that predicts the affine transformations s and t contains 2 fully connected layers with 1024 neurons

and ReLU activation functions. Please see supplemental for more details about normalizing flows.

Training Details. The normalizing flow is pretrained separately from the lifting network for 100 epochs with a batch size of 256 samples. We use the Adam optimizer with an initial learning rate of 10^{-4} and a weight decay of 10^{-5}. For the pretraining of the normalizing flow we divided the learning rate by 10 after 10, 20, and 30 epochs.

The full loss function is

$$\mathcal{L} = \mathcal{L}_{\text{NF}} + 50\mathcal{L}_{\text{bone}} + \mathcal{L}_{\text{base}}. \tag{13}$$

When training the lifting network we use an initial learning rate of $2 \cdot 10^{-4}$ and an exponential scheduling with a decay of 0.95 every epoch for a total number of 100 epochs. Both, the pretraining of the normalizing flow and the training of the lifting network, take approximately 6 hours on an NVIDIA P100 Pascal.

4. Experiments

We perform experiments on the well-known benchmark datasets Human3.6M [15], MPI-INF-3DHP [31] and 3DPW [55]. For the Human3.6M dataset, we follow standard protocols and evaluate on every 64th frame of the test set.

Metrics For the evaluation on Human3.6M we calculate the *mean per joint position error* (MPJPE), i.e. the mean Euclidean distance between the reconstructed and the ground truth joint coordinates. Since an unsupervised setting does not contain metric data we scale the reconstructed 3D pose to match ground truth, commonly known as N-MPJPE [46]. The second common protocol first employs a Procrustes alignment (includes scaling) between the poses before calculating the MPJPE, also known as PA-MPJPE. For 3DHP we report the *Percentage of Correct Keypoints* (PCK) and the corresponding area under curve, scale-normalized as mentioned above, which we call *N-PCK*. It is the percentage of predicted joints that are within a distance of $150mm$ or lower to their corresponding ground truth joint. Additionally, we evaluate the Correct Poses Score (CPS) recently proposed by Wandt et al. [57]. Unlike the PCK, the CPS classifies a pose as correct if all joints of the pose are correctly estimated. To be independent of a threshold value, the CPS calculates the area under the curve in a range from $0mm$ to $300mm$.

4.1. Results in Controlled Conditions

To show the performance of our approach in a fair comparison to others, we start with using the 2D poses given by the dataset. This allows for a fair comparison since it does not rely on the performance of pretrained 2D detectors that vary from method to method. Table 1 presents results for the benchmark dataset Human3.6M with different types of supervision. All shown results use the same 2D input data. We outperform state-of-the-art [63] in unsupervised pose estimation in PA-MPJPE by 12.6%. Notably, we even slightly improve on the PA-MPJPE of the fully supervised method of Martinez *et al*. [30]. We achieve comparable performance to weakly supervised methods and approaches using multi-view supervision in the N-MPJPE metric. Note that [63] uses a prior for the scale during training and therefore directly calculate the MPJPE. However, we outperform them even when they apply the ground truth scale (PA-MPJPE: 39.7) during training. For the Human3.6M dataset we obtain a CPS of 196.1. Table 2 and Table 3 shows results for the MPI-INF-3DHP and 3DPW (in Train-Test-mode) datasets, respectively. On the 3DHP dataset, we only found two other methods that use the 2D poses provided by the dataset and on the 3DPW we found no comparable method. To create training data for the 3DPW dataset we reprojected the 3D skeletons to 2D. This step is necessary because of the difference between the provided 2D and 3D data. Note that for practical applications this step is not required. The 3DPW dataset is particularly challenging since its training set comprises only data captured in-the-wild and additionally it is much smaller compared to the other two datasets. The results show that our approach performs well even in challenging conditions. Furthermore, we evaluate the average distance of the predicted and ground truth elevation angle which is $3.0°$ for Human3.6M and $0.4°$ for MPI-INF-3DHP, respectively.

Table 1. Evaluation results for the Human3.6M dataset in mm. The bottom section, labeled with *unsupervised*, shows comparable unsupervised methods. Best results are marked in bold. Numbers are taken from the respective papers. The star * indicates using a scale prior from the dataset. The MPJPE for [30] is taken from [63].

Supervision	Method	PA-MPJPE↓	N-MPJPE↓
full	Martinez [30]	37.1	45.5*
weak	3D interpreter [60]	88.6	-
	AIGN [54]	79.0	-
	RepNet [56]	38.2	50.9
	Drover [8]	38.2	-
	Kundu [22]	62.4	-
multi-view	EpipolarPose [20]	47.9	54.9
	Wandt [57]	51.4	65.9
unsupervised	Chen [4]	58.0	-
	[4] reimplemented by [63]	46.0	-
	Yu [63] (temporal)	42.0	85.3*
	Ours	**36.7**	64.0

Figure 3 shows subjective results for both datasets. On the left side are reconstructions with a low PA-MPJPE and visually plausible 3D skeletons. Even poses that rarely occur in the training set are reconstructed correctly, e.g., sitting on the floor with crossed legs. The right column shows occasional failure cases, with a PA-MPJPE over 200mm. Typical failure cases are: limbs are rotated in the wrong di-

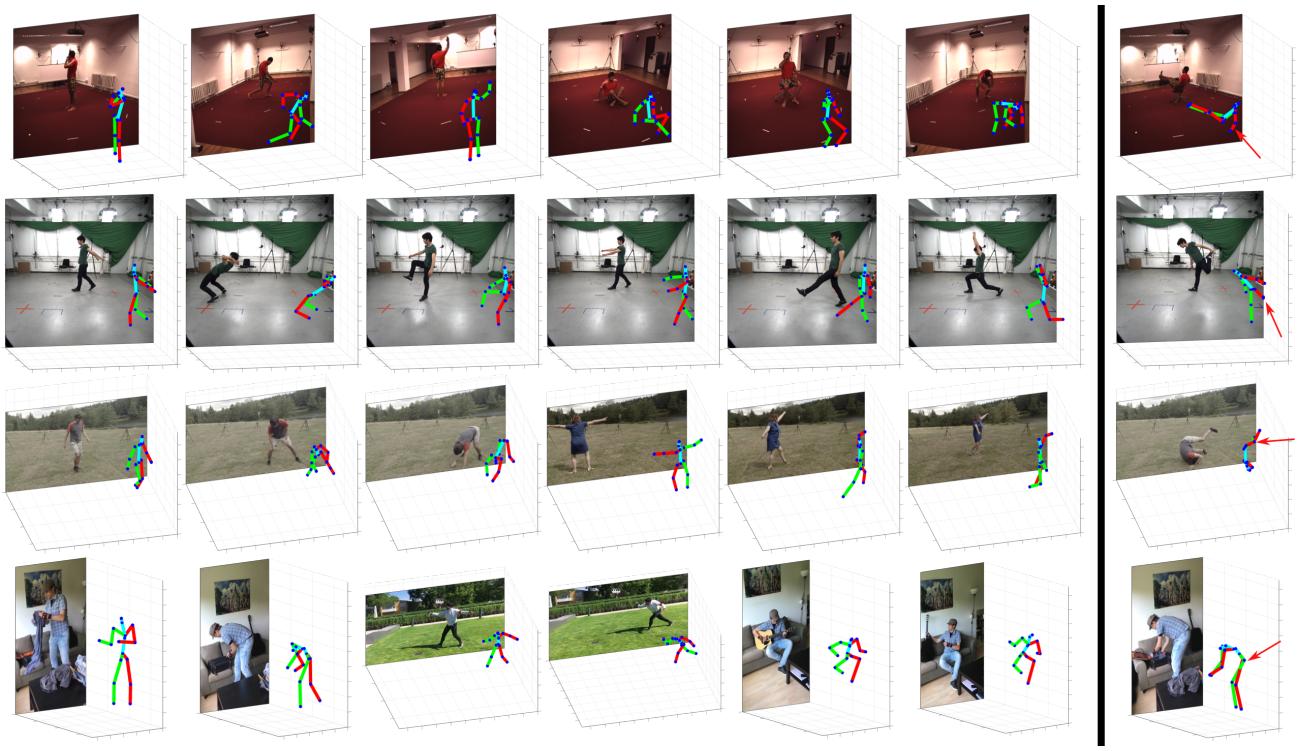

Figure 3. Subjective results of our method for the Human3.6M dataset (top row), the 3DHP dataset (middle two rows), and the 3DPW dataset (last row). The last column shows failure cases.

Table 2. Evaluation results for the MPI-INF-3DHP dataset. The bottom section, labeled with *unsupervised*, shows methods that can solve our setting. Numbers are taken from [63]. A star * indicates an unknown normalization.

Supervision	Method	PA-MPJPE↓	N-PCK↑	AUC↑
weak	Kundu [22]	93.9	84.6	60.8
unsupervised	Yu [63]	-	86.2*	51.7*
	Ours	**54.0**	86.0	50.1

Table 3. Evaluation results for the 3DPW dataset. Results with * do not use ground truth input data.

	Method	PA-MPJPE↓	N-MPJPE↓	N-PCK↑	AUC↑	CPS↑
supervised	Kocabas [19]*	53.2	-	-	-	-
	Lin [28]*	45.6	-	-	-	-
	Li [24]*	48.8	-	-	-	-
	Kocabas [18]*	46.5	-	-	-	-
unsupervised	Ours	64.1	93.0	81.5	51.5	120.3

rection (first and third row) and limb ordering (second and fourth row).

4.2. Results in Practical Conditions

In practice, where only images are available, we use an off-the-shelf 2D pose detector. To be directly comparable to our closest competitor we use the same 2D detections produced by Cascaded Pyramid Networks [5] that are provided by the authors of VideoPose3D [40, 41]. Table 4 shows our results when testing on the predicted 2D poses. We outperform comparable unsupervised methods even though [3, 63] both use temporal information.

Table 4. Results for unsupervised methods of the Human3.6M dataset when using 2D pose predictions. The star * indicates using a scale prior from the dataset instead of applying normalization via N-MPJPE at the inference stage.

	PA-MPJPE↓	N-MPJPE↓	CPS↑
Kundu [22]	62.4	-	-
Kundu [23]	63.8	-	-
Chen [4]	68.0	-	-
Yu [63]	52.3	92.4*	-
Ours	**50.2**	74.4	165.3

4.3. Correlation Between Predicted 3D Poses and Likelihood of Projections

A benefit of our novel normalizing flow formulation is that it can also be used during testing to evaluate the likelihood of the predicted 3D poses. For practical applications, this can be an important value to assess the reliability of the predicted poses for downstream tasks. We apply the normalizing flow to compute the negative log-likelihood of predicted poses. For the reprojection log-likelihood we randomly sample 100 rotations from the distribution learned during the training stage which is then averaged over all

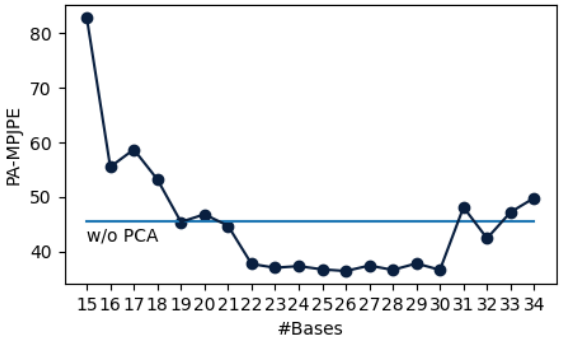

Figure 4. PA-MPJPE for different numbers of PCA bases. Between 22 and 30 PCA bases appears to be the ideal range.

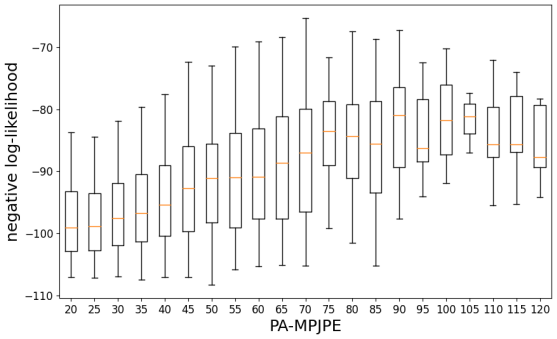

Figure 5. Correlation between the PA-MPJPE and the negative log-likelihood assigned to a set of projections of the predicted 3D pose. As desired, poses with a low 3D error have a low negative log-likelihood and vice versa for high 3D errors. Errors are given in millimeters.

Table 5. Ablation study with different loss terms on the Human3.6M dataset.

Configuration	PA-MPJPE	MPJPE
base ($\mathcal{L}_{NF}$ + $\mathcal{L}_{3D}$ + $\mathcal{L}_{def}$ + $\mathcal{L}_{2D}$)	77.9	135.0
base + $\mathcal{L}_{bone}$	48.1	83.8
base + $\mathcal{L}_{bone}$ + elevation	45.5	73.9
base + $\mathcal{L}_{bone}$ + PCA	43.1	83.8
Ours (base + $\mathcal{L}_{bone}$ + PCA + elevation)	36.7	64.0

rotations. The elevation distribution is estimated over the whole training set. A box plot of the results is shown in Fig. 5. We show bins in 5mm steps from 20-120mm while the 120mm bin includes 120mm and above. As expected, in many cases, the likelihood correlates with the 3D reconstruction error.

4.4. Ablation Studies

We perform several experiments with different configurations of our approach on the Human3.6M dataset by training the lifting network with each of them separately. Additionally, we train the normalizing flow directly on the 2D poses (i.e. no PCA). The results in Table 5 show that each of our contributions is important to achieve the best performance. Note that using the bone lengths prior together with either PCA or elevation alone outperforms [4] and its improved reimplementation by [63]. Without the PCA we achieve an PA-MPJPE of 45.5mm which shows the importance of projecting to the PCA space before training the normalizing flow. Adding our novel elevation prediction improves the results by almost 15%.

Since the PCA is an important part of achieving an acceptable performance, we evaluate the impact of the number of PCA bases. Fig. 4 shows the results. Projecting to a PCA space smaller than 15 bases removes important information from the reprojected 2D poses and results in errors above 100mm. For visualization purposes we only visualize the error for more than 15 bases. The best performance lies between 22 and 30 bases that cover between 99.6% and 99.9% of the variance in the training set. The increase at 31 bases shows that the normalizing flow struggles to learn the probability density when the input dimension is too large.

5. Limitations

The only requirement for our approach is a set of 2D annotations which can be obtained by crowd sourcing 2D joint annotations. This is the main limitation. More specifically, our method requires similar poses seen from different angles. While we compensate for one of these aspects, the elevation angle, natural assumptions on the shape of the distribution of azimuth angles are hard or even impossible to make. Additionally, poses that appear visually correct from all angles can still be implausible in 3D space which is a general problem in monocular human pose estimation. In future work we plan to mitigate these problems by learning 3D pose priors and conditional distributions over full camera rotations jointly.

6. Conclusion

We propose an unsupervised approach that learns to estimate a 3D human pose only from 2D annotations. While previous approaches utilize a predefined prior on the camera distribution of the training set we find that learning this distribution significantly improves the results. Additionally, we utilize normalizing flows to learn a 3D pose prior over random projections of the 3D poses. Moreover, our formulation allows us to calculate the likelihood of reconstructed 3D pose at test time which provides valuable information. Since we observed that directly using the normalizing flow as prior leads to unstable training of the lifting network we additionally propose to first project the 2D poses to a low dimensional subspace.

References

[1] Benjamin Biggs, Sébastien Erhardt, Hanbyul Joo, Benjamin Graham, Andrea Vedaldi, and David Novotny. 3D multibodies: Fitting sets of plausible 3D models to ambiguous image data. In *NeurIPS*, 2020. 3

[2] Ching-Hang Chen and Deva Ramanan. 3d human pose estimation= 2d pose estimation+ matching. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7035–7043, 2017. 2

[3] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M. Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[4] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5714–5724, 2019. 1, 3, 4, 5, 6, 7, 8

[5] 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 conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 7

[6] Zhihua Chen, Xiaoli Liu, Bing Sheng, and Ping Li. Garnet: Graph attention residual networks based on adversarial learning for 3d human pose estimation. In *Advances in Computer Graphics*, pages 276–287, Cham, 2020. Springer International Publishing. 2

[7] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real NVP. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 3

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

[9] Yu Du, Yongkang Wong, Yonghao Liu, Feilin Han, Yilin Gui, Zhen Wang, Mohan Kankanhalli, and Weidong Geng. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *European Conference on Computer Vision (ECCV)*, pages 20–36. Springer, 2016. 2

[10] Haoshu Fang, Yuanlu Xu, Wenguan Wang, Xiaobai Liu, and Song-Chun Zhu. Learning pose grammar to encode human body configuration for 3d pose estimation. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6821–6828. AAAI Press, 2018. 2

[11] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *International Conference on Neural Information Processing Systems (NIPS)*, NIPS'14, pages 2672–2680. MIT Press, 2014. 1

[12] Julian Habekost, Takaaki Shiratori, Yuting Ye, and Taku Komura. Learning 3d global human motion estimation from unpaired, disjoint datasets. 2020. 2

[13] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[14] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Jonathan Eisenmann, Matthew Fisher, Emiliano Gambaretto, Sunil Hadap, and Jean-François Lalonde. A perceptual measure for deep single image camera calibration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2354–2363, 2018. 2

[15] 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 (TPAMI)*, 36(7):1325–1339, 2014. 6

[16] 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 (CVPR)*, June 2020. 3

[17] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[18] Muhammed Kocabas, Chun-Hao P. Huang, Otmar Hilliges, and Michael J. Black. PARE: Part attention regressor for 3D human body estimation. In *Proceedings International Conference on Computer Vision (ICCV)*, pages 11127–11137. IEEE, Oct. 2021. 7

[19] Muhammed Kocabas, Chun-Hao P. Huang, Joachim Tesch, Lea Müller, Otmar Hilliges, and Michael J. Black. SPEC: Seeing people in the wild with an estimated camera. In *Proc. International Conference on Computer Vision (ICCV)*, pages 11035–11045, Oct. 2021. 2, 7

[20] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3, 6

[21] Nikos Kolotouros, Georgios Pavlakos, Dinesh Jayaraman, and Kostas Daniilidis. Probabilistic modeling for human mesh recovery. In *ICCV*, 2021. 3

[22] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 7

[23] Jogendra Nath Kundu, Siddharth Seth, Rahul M. V., Mugalodi Rakesh, Venkatesh Babu Radhakrishnan, and Anirban Chakraborty. Kinematic-structure-preserved representation for unsupervised 3d human pose estimation. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI*, pages 11312–11319. AAAI Press, 2020. 7

[24] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3383–3393, 2021. 7

[25] Sijin Li and Antoni B. Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *Asian Conference on Computer Vision (ACCV)*, volume 9004, pages 332–347, Germany, 11 2014. Springer Verlag. 2

[26] Shichao Li, Lei Ke, Kevin Pratama, Yu-Wing Tai, Chi-Keung Tang, and Kwang-Ting Cheng. Cascaded deep monocular 3d human pose estimation with evolutionary training data. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[27] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, ICCV '15, pages 2848–2856, Washington, DC, USA, 2015. IEEE Computer Society. 2

[28] Kevin Lin, Lijuan Wang, and Zicheng Liu. Mesh graphormer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12939–12948, 2021. 7

[29] Chenxu Luo, Xiao Chu, and Alan L. Yuille. Orinet: A fully convolutional network for 3d human pose estimation. In *British Machine Vision Conference (BMVC)*, page 92, 2018. 2

[30] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3d human pose estimation. In *International Conference on Computer Vision (ICCV)*, 2017. 2, 5, 6

[31] 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 *International Conference on 3D Vision (3DV)*. IEEE, 2017. 2, 6

[32] Dushyant Mehta, Oleksandr Sotnychenko, Franziska Mueller, Weipeng Xu, Mohamed Elgharib, Pascal Fua, Hans-Peter Seidel, Helge Rhodin, Gerard Pons-Moll, and Christian Theobalt. XNect: Real-time multi-person 3D motion capture with a single RGB camera. volume 39, 2020. 2

[33] 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. In *ACM Transactions on Graphics*, volume 36, 7 2017. 2

[34] Rahul Mitra, Nitesh B Gundavarapu, Abhishek Sharma, and Arjun Jain. Multiview-consistent semi-supervised learning for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6907–6916, 2020. 2

[35] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[36] Qiang Nie, Ziwei Liu, and Yunhui Liu. Unsupervised 3d human pose representation with viewpoint and pose disentanglement. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[37] 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 (ECCV)*, pages 156–169, 2016. 2

[38] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition*, pages 7307–7316, 2018. 2

[39] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G. Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1263–1272, 2017. 2

[40] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[41] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. Github, 3d human pose estimation in video with temporal convolutions and semi-supervised training, 2021. 7

[42] Alexis Pietak, Siyan Ma, Caroline W Beck, and Mark D Stringer. Fundamental ratios and logarithmic periodicity in human limb bones. *Journal of Anatomy*, 222:526 – 537, 2013. 5

[43] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d human pose estimation. In *European Conference on Computer Vision (ECCV)*, 2018. 2

[44] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *Conference on Computer Vision and Pattern Recognition*, pages 7703–7713, 2019. 2

[45] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV*, 2018. 2

[46] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8437–8446, 2018. 2, 6

[47] Guillaume Rochette, Chris Russell, and Richard Bowden. Weakly-supervised 3d pose estimation from a single image using multi-view consistency. *BMVC*, 2019. 3

[48] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-Classification-Regression for Human Pose. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1216–1224, Honolulu, United States, July 2017. IEEE. 2

[49] Jennifer J. Sun, Jiaping Zhao, Liang-Chieh Chen, Florian Schroff, Hartwig Adam, and Ting Liu. View-invariant probabilistic embedding for human pose. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[50] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *International Conference on Computer Vision*, pages 2602–2611, 2017. 2

[51] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *European Conference on Computer Vision (ECCV)*, pages 529–545, 2018. 2

[52] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3d human pose with deep neural networks. In *British Machine Vision Conference (BMVC)*, 2016. 2

[53] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[54] Hsiao-Yu F. Tung, Adam W. Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *International Conference on Computer Vision (ICCV)*, pages 4364–4372, 2017. 6

[55] Timo von Marcard, Roberto Henschel, Michael J. Black, Bodo Rosenhahn, and Gerard Pons-Moll. Recovering accurate 3D human pose in the wild using IMUs and a moving camera. In *European Conference on Computer Vision (ECCV)*, volume Lecture Notes in Computer Science, vol 11214, pages 614–631. Springer, Cham, Sept. 2018. 6

[56] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 6

[57] Bastian Wandt, Marco Rudolph, Petrissa Zell, Helge Rhodin, and Bodo Rosenhahn. Canonpose: Self-supervised monocular 3d human pose estimation in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 6

[58] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[59] Tom Wehrbein, Marco Rudolph, Bodo Rosenhahn, and Bastian Wandt. Probabilistic monocular 3d human pose estimation with normalizing flows. In *International Conference on Computer Vision (ICCV)*, Oct. 2021. 3

[60] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *European Conference on Computer Vision (ECCV)*, 2016. 6

[61] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[62] Jingwei Xu, Zhenbo Yu, Bingbing Ni, Jiancheng Yang, Xiaokang Yang, and Wenjun Zhang. Deep kinematics analysis for monocular 3d human pose estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[63] 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 (ICCV)*, pages 8651–8660, October 2021. 1, 3, 4, 5, 6, 7, 8

[64] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Weakly supervised 3d human pose and shape reconstruction with normalizing flows. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 465–481, Cham, 2020. Springer International Publishing. 2, 3

[65] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *International Conference on Computer Vision*, pages 2344–2353, 2019. 2