Weakly-Supervised 3D Human Pose Learning via Multi-view Images in the Wild

Umar Iqbal Pavlo Molchanov Jan Kautz
NVIDIA
{uiqbal, pmolchanov, jkautz}@nvidia.com

Abstract

One major challenge for monocular 3D human pose estimation in-the-wild is the acquisition of training data that contains unconstrained images annotated with accurate 3D poses. In this paper, we address this challenge by proposing a weakly-supervised approach that does not require 3D annotations and learns to estimate 3D poses from unlabeled multi-view data, which can be acquired easily in in-the-wild environments. We propose a novel end-to-end learning framework that enables weakly-supervised training using multi-view consistency. Since multi-view consistency is prone to degenerated solutions, we adopt a 2.5D pose representation and propose a novel objective function that can only be minimized when the predictions of the trained model are consistent and plausible across all camera views. We evaluate our proposed approach on two large scale datasets (Human3.6M and MPII-INF-3DHP) where it achieves state-of-the-art performance among semi-/weakly-supervised methods.

1. Introduction

Learning to estimate 3D body pose from a single RGB image is of great interest for many practical applications. The state-of-the-art methods [6, 16, 17, 28, 32, 39–41, 52, 53] in this area use images annotated with 3D poses and train deep neural networks to directly regress 3D pose from images. While the performance of these methods has improved significantly, their applicability in in-the-wild environments has been limited due to the lack of training data with ample diversity. The commonly used training datasets such as Human3.6M [10], and MPII-INF-3DHP [22] are collected in controlled indoor settings using sophisticated multi-camera motion capture systems. While scaling such systems to unconstrained outdoor environments is impractical, manual annotations are difficult to obtain and prone to errors. Therefore, current methods resort to existing training data and try to improve the generalizabilty of trained models by incorporating additional weak supervision in form of various 2D annotations for in-the-wild images [27, 39, 52]. While 2D annotations can be obtained easily, they do not provide sufficient information about the 3D body pose, especially when the body joints are foreshortened or occluded. Therefore, these methods rely heavily on the ground-truth 3D annotations, in particular, for depth predictions.

Instead of using 3D annotations, in this work, we propose to use unlabeled multi-view data for training. We assume this data to be without extrinsic camera calibration. Hence, it can be collected very easily in any in-the-wild setting. In contrast to 2D annotations, using multi-view data for training has several obvious advantages *e.g.*, ambiguities arising due to body joint occlusions as well as foreshortening or motion blur can be resolved by utilizing information from other views. There have been only few works [14, 29, 33, 34] that utilize multi-view data to learn monocular 3D pose estimation models. While the approaches [29, 33] need extrinsic camera calibration, [33, 34] require at least some part of their training data to be labelled with ground-truth 3D poses. Both of these requirements are, however, very hard to acquire for unconstrained data, hence, limit the applicability of these methods to controlled indoor settings. In [14], 2D poses obtained from multiple camera views are used to generate pseudo ground-truths for training. However, this method uses a pre-trained pose estimation model which remains fixed during training, meaning 2D pose errors remain unaddressed and can propagate to the generated pseudo ground-truths.

In this work, we present a weakly-supervised approach for monocular 3D pose estimation that does not require any 3D pose annotations at all. For training, we only use a collection of unlabeled multi-view data and an independent collection of images annotated with 2D poses. An overview of the approach can be seen in Fig. 1. Given an RGB image as input, we train the network to predict a 2.5D pose representation [12] from which the 3D pose can be reconstructed in a fully-differentiable way. Given unlabeled multi-view data, we use a multi-view consistency loss which enforces the 3D poses estimated from different views to be consistent up to a rigid transformation. However, naively enforcing multi-view consistency can lead to degenerated solutions. We, therefore, propose a novel objective function which is

constrained such that it can only be minimized when the 3D poses are predicted correctly from all camera views. The proposed approach can be trained in a fully end-to-end manner, it does not require extrinsic camera calibration and is robust to body part occlusions and truncations in the unlabeled multi-view data. Furthermore, it can also improve the 2D pose predictions by exploiting multi-view consistency during training.

We evaluate our approach on two large scale datasets where it outperforms existing methods for semi-/weakly-supervised methods by a large margin. We also show that the MannequinChallenge dataset [18], which provides in-the-wild videos of people in static poses, can be effectively exploited by our proposed method to improve the generalizability of trained models, in particular, when their is a significant domain gap between the training and testing environments.

2. Related Work

We discuss existing methods for monocular 3D human pose estimation with varying degree of supervision.

Fully-supervised methods aim to learn a mapping from 2D information to 3D given pairs of 2D-3D correspondences as supervision. The recent methods in this direction adopt deep neural networks to directly predict 3D poses from images [16, 17, 41, 53]. Training the data hungry neural networks, however, requires large amounts of training images with accurate 3D pose annotations which are very hard to acquire, in particular, in unconstrained scenarios. To this end, the approaches in [5, 35, 45] try to augment the training data using synthetic images, however, still need real data to obtain good performance. More recent methods try to improve the performance by incorporating additional data with weak supervision *i.e.*, 2D pose annotations [6, 28, 32, 39, 40, 52], boolean geometric relationship between body parts [27, 31, 37], action labels [20], and temporal consistency [2]. Adverserial losses during training [50] or testing [44] have also been used to improve the performance of models trained on fully-supervised data.

Other methods alleviate the need of 3D image annotations by directly lifting 2D poses to 3D without using any image information *e.g.*, by learning a regression network from 2D joints to 3D [9, 21, 24] or by searching nearest 3D poses in large databases using 2D projections as the query [3, 11, 31]. Since these methods do not use image information for 3D pose estimation, they are prone to re-projection ambiguities and can also have discrepancies between the 2D and 3D poses.

In contrast, in this work, we present a method that combines the benefits of both paradigms *i.e.*, it estimates 3D pose from an image input, hence, can handle the re-projection ambiguities, but does not require any images with 3D pose annotations.

Semi-supervised methods require only a small subset of training data with 3D annotations and assume no or weak supervision for the rest. The approaches [33, 34, 51] assume that multiple views of the same 2D pose are available and use multi-view constraints for supervision. Closest to our approach in this category is [34] in that it also uses multi-view consistency to supervise the pose estimation model. However, their method is prone to degenerated solutions and its solution space cannot be constrained easily. Consequently, the requirement of images with 3D annotations is inevitable for their approach. In contrast, our method is weakly-supervised. We constrain the solution space of our method such that the 3D poses can be learned without any 3D annotations. In contrast to [34], our approach can easily be applied to in-the-wild scenarios as we will show in our experiments. The approaches [43, 48] use 2D pose annotations and re-projection losses to improve the performance of models pre-trained using synthetic data. In [19], a pre-trained model is iteratively improved by refining its predictions using temporal information and then using them as supervision for next steps. The approach in [30] estimates the 3D poses using a sequence of 2D poses as input and uses a re-projection loss accumulated over the entire sequence for supervision. While all of these methods demonstrate impressive results, their main limiting factor is the need of ground-truth 3D data.

Weakly-supervised methods do not require paired 2D-3D data and only use weak supervision in form of motion-capture data [42], images/videos with 2D annotations [25, 47], collection of 2D poses [4, 7, 46], or multi-view images [14, 29]. Our approach also lies in this paradigm and learns to estimate 3D poses from unlabeled multi-view data. In [42], a probabilistic 3D pose model learned using motion-capture data is integrated into a multi-staged 2D pose estimation model to iteratively refine 2D and 3D pose predictions. The approach [25] uses a re-projection loss to train the pose estimation model using images with only 2D pose annotations. Since re-projection loss alone is insufficient for training, they factorize the problem into the estimation of view-point and shape parameters and provide inductive bias via a canonicalization loss. Similar in spirit, the approaches [4, 7, 46] use collection of 2D poses with re-projection loss for training and use adversarial losses to distinguish between plausible and in-plausible poses. In [47], non-rigid structure from motion is used to learn a 3D pose estimator from videos with 2D pose annotations. The closest to our work are the approaches of [14, 29] in that they also use unlabeled multi-view data for training. The approach of [29], however, requires calibrated camera views that are very hard to acquire in unconstrained environments. The approach [14] estimates 2D poses from multi-view images and reconstructs corresponding 3D pose using Epipolar geometry. The reconstructed poses are then used for

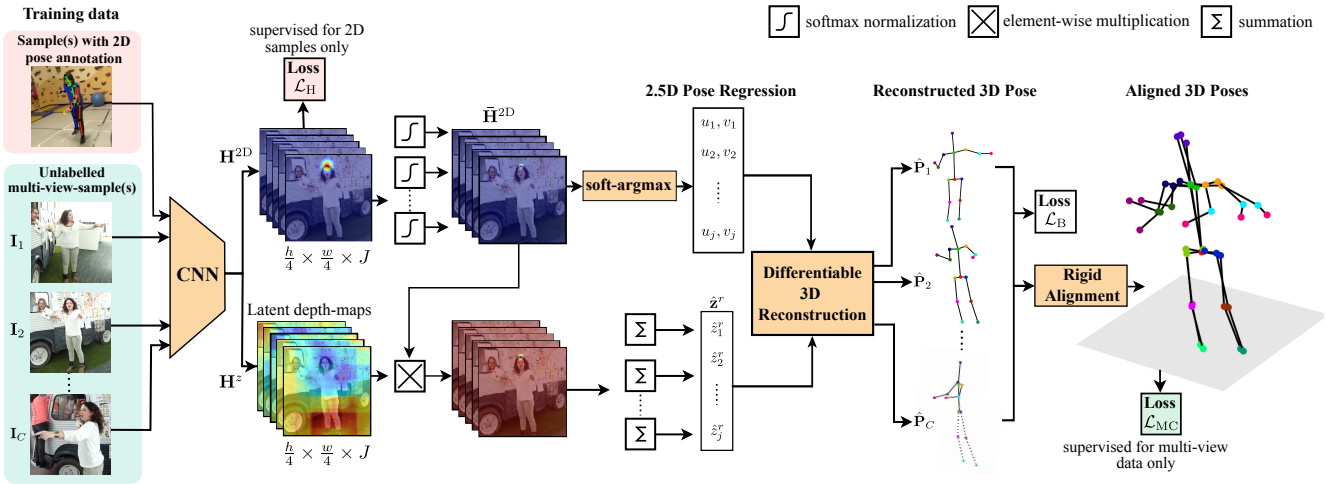

Figure 1. An end-to-end approach for learning 3D pose estimation model without 3D annotations. For training, we only use unlabeled multi-view data along with an independent collection of images with 2D pose annotations. Given an RGB image, the model is trained to generate 2D heatmaps $\mathbf{H}^{\text{2D}}$ and latent depth-maps $\mathbf{H}^{z}$ - shown only for $\mathbf{I}_2$ for simplicity. The 2D heatmaps are converted to 2D pose coordinates using soft-argmax. The relative-depth values $\hat{\mathbf{z}}^{r}$ are obtained by taking channel-wise summation of the multiplication of normalized heatmaps $\bar{\mathbf{H}}^{\text{2D}}$ and latent depth-maps $\mathbf{H}^{z}$. The 3D pose is reconstructed in a fully differentiable manner by exploiting the scale normalization constraint (Sec. 3.1). The images with 2D pose annotation are used for heatmap loss $\mathcal{L}_{\text{H}}$. The 3D supervision is provided via a multi-view consistency loss $\mathcal{L}_{\text{MC}}$ that enforces that the 3D poses generated from different views should be identical up to a rigid transform. Given 2D pose estimates from different views and camera intrinsics, the objective is designed such that the only way for the network to minimize it is to produce correct relative depth values $\hat{\mathbf{z}}^{r}$ (Sec. 3.3). We also enforce a bone-length loss $\mathcal{L}_{\text{B}}$ on each predicted 3D pose to further constrain the search space.

training in a fully-supervised way. The main drawback of this method is that the 3D poses remain fixed throughout the training, and the errors in 3D reconstruction directly propagate to the trained models. This is, particularly, problematic if the multi-view data is captured in challenging outdoor environments where 2D pose estimation may fail easily. In contrast, in this work, we propose an end-to-end learning framework which is robust to challenges posed by the data captured in in-the-wild scenarios. It is trained using a novel objective function which can simultaneously optimize for 2D and 3D poses. In contrast to [14], our approach can also improve 2D predictions using unlabeled multi-view data. We evaluate our approach on two challenging datasets where it outperforms existing methods for semi-/weakly-supervised learning by a large margin.

3. Method

Our goal is to train a convolutional neural network $\mathcal{F}(\mathbf{I}, \theta)$ parameterized by weights θ that, given an RGB image $\mathbf{I}$ as input, estimates the 3D body pose $\mathbf{P} = \{\mathbf{p}_j\}_{j \in J}$ consisting of 3D locations $\mathbf{p}_j = (x_j, y_j, z_j) \in \mathbb{R}^3$ of J body joints with respect to the camera.

We do not assume any training data with paired 2D-3D annotations and learn the parameters θ of the network in a weakly-supervised way using unlabeled multi-view images and an independent collection of images with 2D pose annotations. To this end, we build on the 2.5D pose representation of [12] for hand pose estimation and extend it to human body. This 2.5D pose representation has several key features that allow us to exploit multi-view information and devise loss functions for weakly-supervised training.

In the following, we first recap the 2.5D pose representation (Sec. 3.1) and the approach to reconstruct absolute 3D pose from it (Sec. 3.1.1). We then describe a fully-supervised approach to regress the 2.5D pose using a convolutional neural network (Sec.3.2) followed by our proposed method for weakly-supervised training in Sec. 3.3. An overview of the proposed approach can be seen in Fig. 1.

3.1. 2.5D Pose Representation

Many existing methods [27, 39, 40] for 3D body pose estimation adopt a 2.5D pose representation $\mathbf{P}^{\text{2.5D}} = \{\mathbf{p}_j^{\text{2.5D}} = (u_j, v_j, z_j^r)\}_{j \in J}$ where u_j and v_j are the 2D projection of the body joint j on a camera plane and $z_j^r = z_{\text{root}} - z_j$ represents its metric depth with respect to the root joint. This decomposition of 3D joint locations into their 2D projection and relative depth has the advantage that additional supervision from in-the-wild images with only 2D pose annotations can be used for better generalization of the trained models. However, this representation does not account for scale ambiguity present in the image which might lead to ambiguities in predictions.

The 2.5D representation of [12], however, differs from the rest in terms of scale normalization of 3D poses. Specifically, they scale normalize the 3D pose $\mathbf{P}$ such that a spe-

cific pair of body joints has a unit distance:

$$\hat{\mathbf{P}} = \frac{\mathbf{P}}{s}, \tag{1}$$

where $s = \|\mathbf{p}_k - \mathbf{p}_l\|_2$ is estimated independently for each pose. The pair (k, l) corresponds to the indices of the joints used for scale normalization. The resulting scale normalized 2.5D pose representation $\hat{\mathbf{p}}_j^{2.5\mathrm{D}} = (u_j, v_j, \hat{z}_j^r)$ is agnostic to the scale of the person. This not only makes it easier to be estimated from cropped RGB images, but also allows to reconstruct the absolute 3D pose of the person up to a scaling factor in a fully differentiable manner as described next.

3.1.1 Differentiable 3D Reconstruction

Given the 2.5D pose $\hat{\mathbf{P}}^{2.5\mathrm{D}}$, we need to find the depth $\hat{z}_{\mathrm{root}}$ of the root joint to reconstruct the scale normalized 3D locations $\hat{\mathbf{P}}$ of body joints using perspective projection:

$$\hat{\mathbf{p}}_j = \hat{z}_j \mathbf{K}^{-1} \begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix} = (\hat{z}_{\mathrm{root}} + \hat{z}_j^r)\mathbf{K}^{-1} \begin{bmatrix} u_j \\ v_j \\ 1 \end{bmatrix}. \tag{2}$$

The value of $\hat{z}_{\mathrm{root}}$ can be calculated via the scale normalization constraint:

$$(\hat{x}_k - \hat{x}_l)^2 + (\hat{y}_k - \hat{y}_l)^2 + (\hat{z}_k - \hat{z}_l)^2 = 1, \tag{3}$$

which leads to an analytical solution as derived in [12]. Since all operations for 3D reconstruction are differentiable, we can devise loss functions that directly operate on the reconstructed 3D poses.

In the rest of this paper, we will use the scale normalized 2.5D pose representation. We use the distance between the neck and pelvis joints to calculate the scaling factor s.

3.2. 2.5D Pose Regression

Since the 3D pose can be reconstructed analytically from 2.5D pose, we train the network to predict 2.5D pose and implement 3D reconstruction as an additional parameter-free layer. To this end, we adopt the 2.5D heatmap regression approach of [12]. Specifically, given an RGB image as input, the network produces $2J$ channels as output with J channels for 2D heatmaps ($\mathbf{H}^{2\mathrm{D}}$) while the remaining J channels are regarded as latent depth maps $\mathbf{H}^z$. The 2D heatmaps are converted to 2D pose coordinates (u_j, v_j) by first normalizing them using spatial $\mathrm{softmax}$, *i.e.*, $\bar{\mathbf{H}}_j^{2\mathrm{D}} = \mathrm{softmax}(\mathbf{H}_j^{2\mathrm{D}}, \lambda)$, and then using the $\mathrm{soft\text{-}argmax}$ operation:

$$u_j = \sum_{u,v \in \mathcal{U}} u \cdot \bar{\mathbf{H}}_j^{2\mathrm{D}}(u, v); \quad v_j = \sum_{u,v \in \mathcal{U}} v \cdot \bar{\mathbf{H}}_j^{2\mathrm{D}}(u, v), \tag{4}$$

where $\mathcal{U}$ is a 2D grid sampled according to the effective stride size of the network, and λ is a constant that controls the temperature of the normalized heatmaps.

The relative scale normalized depth value $\hat{z}_j^r$ for each joint can then be obtained as the summation of the element-wise multiplication of $\bar{\mathbf{H}}_j^{2\mathrm{D}}$ and latent depth maps $\mathbf{H}_j^z$:

$$\hat{z}_j^r = \sum_{u,v} \bar{\mathbf{H}}_j^{2\mathrm{D}} \odot \mathbf{H}_j^z. \tag{5}$$

Given the 2D pose coordinates $\{(u_j, v_j)\}_{j \in J}$, relative depths $\hat{\mathbf{z}}^r = \{\hat{z}_j^r\}_{j \in J}$ and intrinsic camera parameters $\mathbf{K}$, the 3D pose can be reconstructed as explained in Sec. 3.1.1.

In the fully-supervised (FS) setting, the network can be trained using the following loss function:

$$\mathcal{L}_{\mathrm{FS}} = \mathcal{L}_{\mathrm{H}}(\mathbf{H}^{2\mathrm{D}}, \mathbf{H}_{\mathrm{gt}}^{2\mathrm{D}}) + \psi \mathcal{L}_{\mathbf{z}}(\hat{\mathbf{z}}^r, \hat{\mathbf{z}}_{\mathrm{gt}}^r), \tag{6}$$

where $\mathbf{H}_{\mathrm{gt}}^{2\mathrm{D}}$ and $\hat{\mathbf{z}}_{\mathrm{gt}}^r$ are the ground-truth 2D heatmaps and ground-truth scale-normalized relative depth values, respectively. We use mean squared error as the loss functions $\mathcal{L}_{\mathrm{H}}(\cdot)$ and $\mathcal{L}_{\mathbf{z}}(\cdot)$.

We make one modification to the original loss to better learn the confidence scores of predictions. Specifically, in contrast to [12], we do not learn 2D heatmaps in a latent way. Instead, we chose to explicitly supervise the 2D heatmap predictions via ground-truth heatmaps with Gaussian distributions at the true joint locations. We will rely on the confidence scores to devise a weakly-supervised loss that is robust to uncertainties in 2D pose estimates, as described in the following section.

3.3. Weakly-Supervised Training

We describe our proposed approach for training the regression network in a weakly-supervised way without any 3D annotations. For training, we assume a set $\mathbf{M} = \{\{\mathbf{I}_c^n\}_{c \in C_n}\}_{n \in N}$ of N samples, with the nth sample consisting of C_n camera views of a person in same body pose. The multi-view images can be taken at the same time using multiple cameras, or using a single camera assuming a static body pose over time. We do not assume knowledge of extrinsic camera parameters. Additionally, we use an independent set of images annotated only with 2D poses which is available abundantly or can be annotated by people even for in-the-wild data. For training, we optimize the following weakly-supervised (WS) loss function:

$$\mathcal{L}_{\mathrm{WS}} = \mathcal{L}_{\mathrm{H}}(\mathbf{H}^{2\mathrm{D}}, \mathbf{H}_{\mathrm{gt}}^{2\mathrm{D}}) + \alpha \mathcal{L}_{\mathrm{MC}}(\mathbf{M}) + \beta \mathcal{L}_{\mathrm{B}}(\hat{\mathbf{L}}, \hat{\mu}^L), \tag{7}$$

where $\mathcal{L}_{\mathrm{H}}$ is the 2D heatmap loss, $\mathcal{L}_{\mathrm{MC}}$ is the multi-view consistency loss, and $\mathcal{L}_{\mathrm{B}}$ is the limb length loss.

Recall that, given an RGB image, our goal is to estimate the scale normalized 2.5D pose $\hat{\mathbf{P}}^{2.5\mathrm{D}} = \{\mathbf{p}_j^{2.5\mathrm{D}} = (u_j, v_j, \hat{z}_j^r)\}_{j \in J}$ from which we can reconstruct the scale normalized 3D pose $\hat{\mathbf{P}}$ as explained in Sec. 3.1.1. While $\mathcal{L}_{\mathrm{H}}$ provides supervision for 2D pose estimation, the loss $\mathcal{L}_{\mathrm{MC}}$ supervises the relative depth component ($\hat{\mathbf{z}}^r$). The

limb length loss $\mathcal{L}_{\text{B}}$ further ensures that the reconstructed 3D pose $\hat{\mathbf{P}}$ has plausible limb lengths. In the following, we explain these loss functions in more detail.

Heatmap Loss ($\mathcal{L}_{\text{H}}$) measures the difference between the predicted 2D heatmaps $\mathbf{H}^{\text{2D}}$ and ground-truth heatmaps $\mathbf{H}^{\text{2D}}_{\text{gt}}$ with Gaussian distribution at the true joint location. It operates only on images annotated with 2D poses and is assumed to be zero for all other images.

Multi-View Consistency Loss ($\mathcal{L}_{\text{MC}}$) enforces that the 3D pose estimates obtained from different views should be identical up to a rigid transform. Formally, given a multi-view training sample $\mathbf{M} = \{\mathbf{I}_c\}_{c \in C}$ with C camera views, we define the multi-view consistency loss as the weighted sum of the difference between the 3D joint locations across different views after the rigid alignment:

$$\mathcal{L}_{\text{MC}} = \sum_{\substack{c,c' \in C \\ c \neq c'}} \sum_{j \in J} \phi_{j,c}\phi_{j,c'} \cdot d(\hat{\mathbf{p}}_{j,c}, \mathbf{R}_c^{c'}\hat{\mathbf{p}}_{j,c'}), \quad (8)$$

where

$$\phi_{j,c} = \mathbf{H}^{\text{2D}}_{j,c}(u_{j,c}, v_{j,c}) \text{ and } \phi_{j,c'} = \mathbf{H}^{\text{2D}}_{j,c'}(u_{j,c'}, v_{j,c'})$$

are the confidence scores of the jth joint in camera viewpoint $\mathbf{I}_c$ and $\mathbf{I}_{c'}$, respectively. The $\hat{\mathbf{p}}_{j,c}$ and $\hat{\mathbf{p}}_{j,c'}$ are the scale normalized 3D coordinates of the jth joint estimated from viewpoint $\mathbf{I}_c$ and $\mathbf{I}_{c'}$, respectively. $\mathbf{R}_c^{c'} \in \mathbb{R}^{3\times 4}$ is a rigid transformation matrix that best aligns the two 3D poses, and d is the distance metric used to measure the difference between the aligned poses. In this work, we use L_1-norm as the distance metric d. In order to understand the contribution of $\mathcal{L}_{\text{MC}}$ more clearly, we can rewrite the distance term in (8) in terms of the 2.5D pose representation using (2), *i.e.*:

$$d(\hat{\mathbf{p}}_{j,c}, \mathbf{R}_c^{c'}\hat{\mathbf{p}}_{j,c'}) = d((\hat{z}_{\text{root},c} + \hat{z}^r_{j,c})\mathbf{K}_c^{-1}\begin{bmatrix} u_{j,c} \\ v_{j,c} \\ 1 \end{bmatrix}, \mathbf{R}_c^{c'}(\hat{z}_{\text{root},c'} + \hat{z}^r_{j,c'})\mathbf{K}_{c'}^{-1}\begin{bmatrix} u_{j,c'} \\ v_{j,c'} \\ 1 \end{bmatrix}). \quad (9)$$

Let us assume that the 2D coordinates $(u_{j,c}, v_{j,c})$ and $(u_{j,c'}, v_{j,c'})$ are predicted accurately due to the loss $\mathcal{L}_{\text{H}}$ and the camera *intrinsics* $\mathbf{K}_c$ and $\mathbf{K}_{c'}$ are known. For simplicity, let us also assume the ground-truth transformation $\mathbf{R}_c^{c'}$ between the two views is known. Then, the only way for the network to minimize the difference $d(.,.)$ is to predict the correct values for relative depths $\hat{z}^r_{j,c}$ and $\hat{z}^r_{j,c'}$. Hence, the joint optimization of the losses $\mathcal{L}_{\text{H}}$ and $\mathcal{L}_{\text{MC}}$ allows us to learn correct 3D poses using only weak supervision in form of multi-view images and 2D pose annotations. Without the loss $\mathcal{L}_{\text{H}}$ the model can lead to degenerated solutions.

While in many practical scenarios, the transformation matrix $\mathbf{R}_c^{c'}$ can be known a priori via extrinsic calibration, we, however, assume it is not available and estimate it using predicted 3D poses and Procrustes analysis as follows:

$$\mathbf{R}_c^{c'} = \underset{\mathbf{R}}{\text{argmin}} \sum_{j \in J} \phi_{j,c}\phi_{j,c'} \|\hat{\mathbf{p}}_{j,c} - \mathbf{R}\hat{\mathbf{p}}_{j,c'}\|_2^2. \quad (10)$$

During training, we follow [34] and do not back-propagate through the optimization of transformation matrix (10), since it leads to numerical instabilities arising due to singular value decomposition. Note that the gradients from $\mathcal{L}_{\text{MC}}$ not only influence the depth estimates, but also affect heatmap predictions due to the calculation of $\hat{z}_{\text{root}}$ in (3). Therefore, $\mathcal{L}_{\text{MC}}$ can also fix the errors in 2D pose estimates as we will show in our experiments.

Limb Length Loss ($\mathcal{L}_{\text{B}}$) measures the deviation of the limb lengths of predicted 3D pose from the mean bone lengths:

$$\mathcal{L}_{\text{B}} = \sum_{j,j' \in \mathcal{E}} \phi_j\phi_{j'}(\|\hat{\mathbf{p}}_j - \hat{\mathbf{p}}_{j'}\| - \hat{\mu}^L_{j,j'})^2, \quad (11)$$

where $\mathcal{E}$ corresponds to the used kinematic structure of the human body and $\hat{\mu}^L_{j,j'}$ is the scale normalized mean limb length for joint pair (j, j'). Since the limb lengths of all people will be roughly the same after scale normalization (1), this loss ensures that the predicted poses have plausible limb lengths. During training, we found that having a limb length loss leads to faster convergence.

Additional Regularization We found that if a large number of samples in multi-view data have a constant background, the network learns to recognize these images and starts predicting same 2D pose and relative depth values for such images. Interestingly, it predicts correct values for other samples. In order to prevent this, we incorporate an additional regularization loss for such samples. Specifically, we run a pre-trained 2D pose estimation model and generate pseudo ground-truths by selecting joint estimates with confidence score greater than a threshold $\tau = 0.5$. These pseudo ground-truths are then used to enforce the 2D heatmap loss $\mathcal{L}_{\text{H}}$, which prevents the model from predicting degenerated solutions. We generate the pseudo ground-truths once at the beginning of the training and keep them fixed throughout. Specifically, we use the regularization loss for images from Human3.6M [10] and MPII-INF-3DHP [22] that are both recorded in controlled indoor settings. While the regularization may reduce the impact of $\mathcal{L}_{\text{MC}}$ on 2D poses, the gradients from $\mathcal{L}_{\text{MC}}$ will still influence the heatmap predictions of body joints that were not detected with high confidence (see Fig. 2).

4. Experiments

We evaluate our proposed approach for weakly-supervised 3D body pose learning and compare it with the

state-of-the-art methods. Additional training and implementation details can be found in the supplementary material.

4.1. Datasets

We use two large-scale datasets, Human3.6M [10] and MPII-INF-3DHP [22] for evaluation. For weakly-supervised training, we also use the MannequinChallenge dataset [18] and MPII Human Pose dataset [1] . The details of each dataset are as follows.

Human3.6M (H36M) [10] provides images of actors performing a variety of actions from four views. We follow the standard protocol and use five subjects (S1, S5, S6, S7, S8) for training and test on two subjects (S9 and S11).

MPII-INF-3DH (3DHP) [22] provides ground-truth 3D poses obtained using markerless motion-capture system. Following the standard protocol [22], we use five chest height cameras for training. The test-set consists of six sequences with actors performing a variety activities.

MannequinChallenge Dataset (MQC) [18] provides in-the-wild videos of people in static poses while a hand-held camera pans around the scene. The videos do not come with any ground-truth annotations, however, the data is very adequate for our proposed weakly-supervised approach using multi-view consistency. The dataset consists of three splits for training, validation and testing. In this paper, we use $\sim$3300 videos from training and validation set as proposed by [18], but in practice one could download an immense amount of such videos from YouTube (#MannequinChallenge). We will show in our experiments that using these in-the-wild videos during training yields better generalization, in particular, when there is a significant domain gap between the training and testing set. Since the videos can have multiple people inside each frame, they have to be associated across frames to obtain the required multi-view data. To this end, we adopt the pose based tracking approach of [49] and generate person tracklets from each video. For pose estimation, we use a HRNet-w32 [38] model pretrained on MPII Pose dataset [1]. In order to avoid training on noisy data, we discard significantly occluded or truncated people. We do this by discarding all poses that have more than half of the estimated body joints with confidence score lower than a threshold τ=0.5. We also discard poses in which neck or pelvis joints have confidence lower than τ=0.5 since both joints are important for $\hat{z}_{\text{root}}$ reconstruction using (3). Finally, we discard all tracklets with the length lower than 5 frames. This gives us 11,413 multi-view tracklets with 241k images in total. The minimum and maximum length of the tracklets is 5 and 140 frames, respectively.

MPII Pose Dataset (MPII) [1] provides 2D pose annotations for 28k in-the-wild images.

Method	Supervision 2D	3D	MV	Error 2D px	3D mm
FS	H+M	H	-	5.9	55.5
WS + $\mathbf{R}$	H+M	-	H	6.1	57.2
WS	H+M	-	H	6.1	59.3
2d-only	M	-	-	8.9	-
WS + $\mathbf{R}$	M	-	H	8.3	62.3
WS	M	-	H	8.4	69.1
WS	M	-	I	9.0	106.2
WS	M	-	I+Q	9.1	93.6
WS	M	-	H+I+Q	8.4	67.4
WS + $\mathbf{R}$	M	-	H+I+Q	8.4	60.3

Table 1. **Ablative study:** We provide results when different levels of supervision are used to train the proposed weakly-supervised method. FS: Fully-Supervised, WS: Weakly-Supervised, MV: Multi-View, H: H36M, M: MPII, I: 3DHP, Q: MQC. No 3D supervision is used for all experiments except FS.

4.2. Evaluation Metrics

For evaluation on H36M, we follow the standard protocols and use MPJPE (Mean Per Joint Position Error), N-MPJPE (Normalized-MPJPE) and P-MPJPE (Procrustes-aligned MPJPE) for evaluations. MPJPE measures the mean euclidean error between the ground-truth and estimated location of 3D joints after root alignment. While NMPJPE [34] also aligns the scale of the predictions with ground-truths, PMPJPE aligns both the scale and rotations using Procrustes analysis. For evaluation on 3DHP dataset, we follow [22] and also report PCK (Percentage of Correct Keypoints) and Normalized-PCK as defined in [34]. PCK measures the percentage of predicted body joints that lie within the radius of 150mm from their ground-truths. 3DHP evaluation protocol uses 14 joints for evaluation excluding the pelvis joint which is used for alignment of the poses.

4.3. Ablation Studies

Tab. 1 evaluates the impact of different levels of supervision for training with the proposed approach. We use H36M for evaluation. We start with a fully-supervised setting (FS) which uses 2D supervision from H36M and MPII (2D=H+M) datasets and 3D pose supervision from H36M (3D=H). No multi-view (MV) data is used in this case. The fully-supervised model yields a MPJPE of 5.9px and 55.5mm for 2D and 3D pose estimation, respectively. We then remove the 3D supervision and instead train the network using the proposed approach for weakly-supervised learning (WS+$\mathbf{R}$). The MV data is taken from H36M (MV=H). For this experiment, we assume that the 2D pose annotations for MV data are available (2D=H+M) and the camera extrinsics $\mathbf{R}$ are known. This setting is essentially similar to fully-supervised case since the 2D poses from different views can be triangulated using the known $\mathbf{R}$. Train-

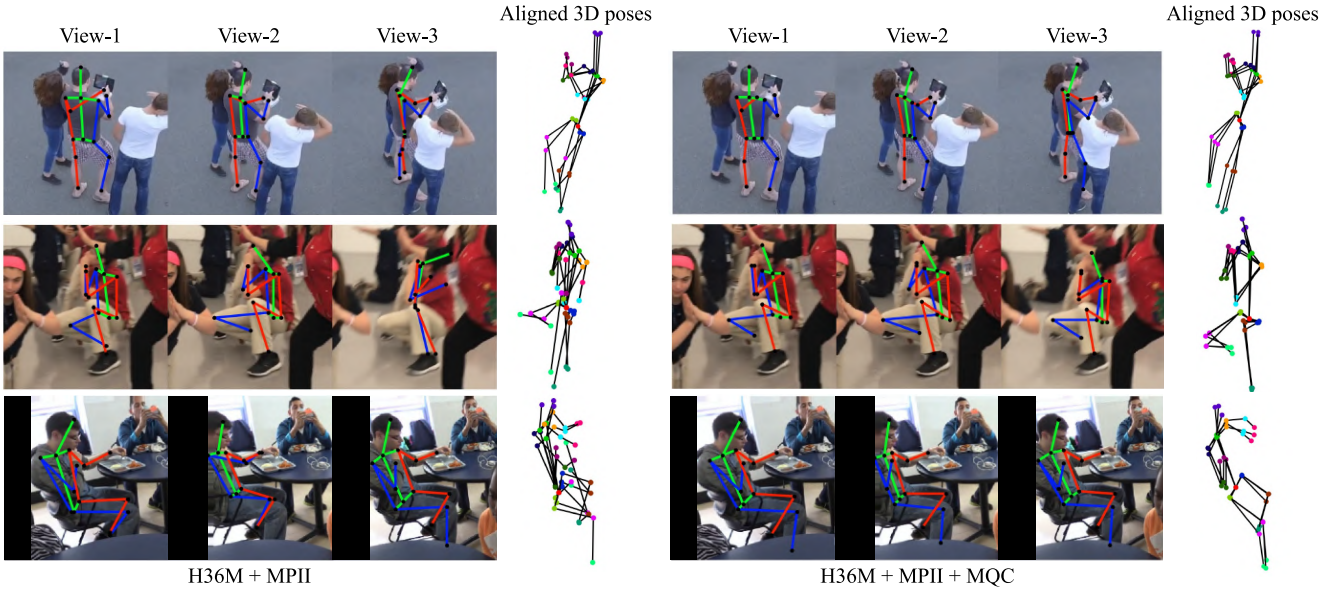

Figure 2. Impact of using MQC dataset. We run the trained models on the tracks taken from MQC dataset and align the estimated 3D poses using (10). Since people in MQC dataset do not move, the aligned poses should be very similar. Adding MQC dataset for training **(right)** yields more consistent 3D pose estimates as compared to when only H36M is used **(left)** for multi-view consistency loss. Note that our proposed approach can also fix the errors in 2D pose estimates in the unlabeled multi-view data.

ing the network under this setting, however, serves as a sanity check that the proposed weakly-supervised approach works as intended which can be confirmed by the obtained 3D pose error of 57.2mm. If $\mathbf{R}$ is unknown (WS) and is obtained from estimated 3D poses using (10), the error increases slightly to 59.3mm.

All of the aforementioned settings assume that the MV data is annotated with 2D poses which is infeasible to collect in large numbers. Therefore, we have designed the proposed method to work with MV data without even 2D annotations. Next, we remove the 2D supervision from MV data and only use MPII dataset for 2D supervision (2D=M). For reference, we also report the error of a 2D-only model trained on MPII dataset which yields a 2D pose error of 8.9px. Training without 2D pose annotations for MV data with and without ground-truth $\mathbf{R}$ yields errors of 62.3mm (WS+$\mathbf{R}$) and 69.1mm (WS), respectively, as compared to 57.2mm and 59.3mm when the 2D pose annotations are available. While using ground-truth $\mathbf{R}$ always yields better performance, for the sake of easier applicability, in the rest of this paper, we assume it to be unknown unless specified otherwise. It is also interesting to note that the 2D pose error decreases from 8.9px to 8.3px when the multi-view consistency loss (8) is used. Some qualitative examples of improvements in 2D poses can be seen in Fig. 2.

We also evaluate the case when the training data is recorded in different settings than testing data. For this, we use 3DHP for training (MV=I) and test on H36M. Since the images of 3DHP are very different from H36M, it leads to a very high error of 106.2mm. Adding the generated training data from MQC dataset (MV=I+Q) significantly reduces the error to 93.6mm which demonstrates the effectiveness of in-the-wild data from MQC. Combining all three datasets (MV=H+I+Q) reduces the error further to 67.4mm as compared to 69.1mm when only H36M dataset was used for training. We also provide the results when ground-truth $\mathbf{R}$ is known (WS+$\mathbf{R}$) for H36M and 3DHP datasets (MV=H+I+Q) which shows a similar behaviour and decreases the error from 62.3mm to 60.3mm.

In our experiments, we found that training only on MQC dataset is not sufficient for convergence and it has to be combined with another dataset which provides multi-view data from more distant viewing angles. This is likely because most videos in MQC dataset do not capture same person from very different viewing angles, whereas datasets such as H36M and 3DHP provide images from cameras with sufficiently large baselines.

4.4. Comparison with the State-of-the-Art

Tab. 2 compares the performance of our proposed method with the state-of-the-art on H36M dataset. We group all approaches in three categories; fully-supervised, semi-supervised, and weakly-supervised, and compare the performance of our method under each category. While fully-supervised methods use complete training set of H36M for 3D supervision, semi-supervised methods use 3D supervision from only one subject (S1) and use other subjects (S5, S6, S7, S9) for weak supervision. Weakly-supervised methods do not use any 3D supervision. Some methods also use ground-truth information during infer-

Methods	MPJPE ↓	NMPJPE ↓	PMPJPE ↓
Fully-Supervised Methods			
Rogez *et al.* [36] (CVPR'17)	87.7	-	71.6
Habibie *et al.* [8] (ICCV'19)	-	65.7	-
Rhodin *et al.* [34] (CVPR'18)	66.8	63.3	51.6
Zhou *et al.* [52] (ICCV'17)	64.9	-	-
Martinez *et al.* [21] (ICCV'17)	62.9	-	47.7
Sun *et al.* [39] (ICCV'17) *	59.6	-	-
Yang *et al.* [50] (CVPR'18)	58.6	-	-
Pavlakos *et al.* [27] (CVPR'18)	56.2	-	-
Sun *et al.* [40] (ECCV'18)*	**49.6**	-	40.6
Kocabas *et al.* [14] (CVPR'19)*	51.8	51.6	45.0
Ours - H - baseline	55.5	51.4	41.5
Ours* - H - baseline	50.2	**49.9**	**36.9**
Ours - H+I+Q	56.1	52.7	45.9
Semi-Supervised Methods - only Subject-1 is used for training			
Rohdin *et al.* [33] (ECCV'18)	131.7	122.6	98.2
Pavlakos *et al.* [26] (ICCV'19)	110.7	97.6	74.5
Li *et al.* [19] (ICCV'19)	88.8	80.1	66.5
Rhodin *et al.* [34] (CVPR'18)	n/a	80.1	65.1
Kocabas *et al.* [14] (CVPR'19)	n/a	67.0	60.2
Ours - H	62.8	59.6	51.4
Ours - H+I+Q	**59.7**	**56.2**	**50.6**
Weakly-Supervised Methods - no 3D supervision			
Pavlakos *et al.* [29] (CVPR'17)	118.4	-	-
Kanzawa *et al.* [13] (CVPR'18)	106.8	-	67.5
Wandt *et al.* [46] (CVPR'19)	89.9	-	-
Tome *et al.* [42] (CVPR'17)	88.4	-	-
Kocabas *et al.* [14] (CVPR'19)	n/a	77.75	70.67
Chen *et al.* [4] (CVPR'19)	-	-	68.0
Drover *et al.* [7] (ECCV-W'18)	-	-	64.6
Kolotouros *et al.* [15] (ICCV'19)	-	-	62.0
Wang *et al.* [47] (ICCV'19)	83.0	-	57.5
Ours - H	69.1	66.3	55.9
Ours - H+I+Q	**67.4**	**64.5**	**54.5**

Table 2. Comparison with the state-of-the-art on H36M dataset. *use ground-truth depth of the root keypoint during inference.

ence [14, 39, 40]. For a fair comparison with those, we also report our performance under the same settings. It is important to note that, many approaches such as [4, 7, 21, 33, 34, 36, 50] estimate root-relative 3D pose. Our approach, on the other hand, estimates absolute 3D poses. While our fully-supervised baseline (Ours-H-baseline) performs better or on-par with the state-of-the-art fully-supervised methods, our proposed approach for weakly-supervised learning significantly outperforms other methods under both semi- and weakly-supervised categories.

For a fair comparison with other methods, we report results of our method under two settings: i) using H36M and MPII dataset for training (Ours-H), and ii) with multi-view data from 3DHP and MQC as additional weak supervision (Ours-H+I+Q). In the fully-supervised case, using additional weak supervision slightly worsens the performance (55.5mm vs 56.1mm) which is not surprising on a dataset like H36M which is heavily biased to indoor data and have training and testing images recorded with a same background. Whereas, our approach, in particular the data from MQC, is devised for in-the wild generalization. The importance of additional multi-view data, however, can be seen evidently in the semi-/weakly-supervised settings where it decreases the error from 62.8mm to 59.7mm and from 69.1mm to 67.4mm, respectively.

Methods	MPJPE↓	NMPJPE↓	PCK↑	NPCK↑
Fully-Supervised Methods				
Mehta *et al.* [23]	-	-	76.6	-
Rohdin *et al.* [34]	n/a	101.5	n/a	78.8
Kocabas *et al.* [14]*	109.0	106.4	77.5	78.1
Ours	**110.8**	**98.9**	**80.2**	**82.3**
Ours*	**99.2**	**97.2**	**83.0**	**83.3**
Semi-Supervised Methods				
Rhodin *et al.* [34]	n/a	121.8	n/a	72.7
Kocabas *et al.* [14]	n/a	119.9	n/a	73.5
Ours	**113.8**	**102.2**	**79.1**	**81.5**
Weakly-Supervised Methods				
Kanazwa *et al.* [13]	169.5	-	59.6	-
Kolotouros *et al.* [15]	124.8	-	66.8	-
Ours	**122.4**	**110.1**	**76.5**	**79.4**
Kocabas *et al.* [14]* + $\mathbf{R}$	126.8	125.7	64.7	71.9
Ours* + $\mathbf{R}$	**109.3**	**107.2**	**79.5**	**80.0**

Table 3. Comparison with the state-of-the-art on 3DHP dataset. *use ground-truth 3D location of the root joint during inference.

Compared to the state-of-the-art method [14] that also uses multi-view information for weak supervision, our method performs significantly better even though the fully-supervised baselines of both approaches perform similar. This demonstrates the effectiveness of our end-to-end training approach and proposed loss functions that are robust to errors in 2D poses. While our weakly-supervised approach does not outperform fully-supervised methods, it performs on-par with many recent fully-supervised approaches.

Tab. 3 compares the performance of our proposed approach with the state-of-the-art on 3DHP dataset. We use our models trained with *Ours-H+I+Q* setting, as described above. We do not use any 3D pose supervision from 3DHP dataset and instead use the same models used for evaluation on H36M dataset. Our proposed approach outperforms all existing methods with large margins under all three categories which also demonstrates the cross dataset generalization of our proposed method.

Some qualitative results of the proposed approach can be seen in the supplementary material.

5. Conclusion

We have presented a weakly-supervised approach for 3D human pose estimation in the wild. Our proposed approach does not require any 3D annotations and can learn to estimate 3D poses from unlabeled multi-view data. This is made possible by a novel end-to-end learning framework and a novel objective function which is optimized to predict consistent 3D poses across different camera views. The proposed approach is very practical since the required training data can be collected very easily in in-the-wild scenarios. We demonstrated state-of-the-art performance on two challenging datasets.

Acknowledgments: We are thankful to Kihwan Kim and Adrian Spurr for helpful discussions.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2D human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 6

[2] Anurag Arnab, Carl Doersch, and Andrew Zisserman. Exploiting temporal context for 3d human pose estimation in the wild. In *CVPR*, 2019. 2

[3] Ching-Hang Chen and Deva Ramanan. 3D human pose estimation = 2D pose estimation + matching. In *CVPR*, 2017. 2

[4] 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 *CVPR*, 2019. 2, 8

[5] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen. Synthesizing training images for boosting human 3d pose estimation. In *3DV*, 2016. 2

[6] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *ECCV*, 2018. 1, 2

[7] Dylan Drover, Rohith M. V, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3d pose be learned from 2d projections alone? In *ECCV Workshops*, 2018. 2, 8

[8] 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 *ICCV*, 2019. 8

[9] Mir Rayat Imtiaz Hossain and James J. Little. Exploiting temporal information for 3d pose estimation. In *ECCV*, 2018. 2

[10] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *TPAMI*, 36(7):1325–1339, 2014. 1, 5, 6

[11] Umar Iqbal, Andreas Doering, Hashim Yasin, Björn Krüger, Andreas Weber, and Juergen Gall. A dual-source approach for 3D human pose estimation in single images. *CVIU*, 2018. 2

[12] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. Hand pose estimation via 2.5D latent heatmap regression. In *ECCV*, 2018. 1, 3, 4

[13] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 8

[14] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *CVPR*, 2019. 1, 2, 3, 8

[15] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 8

[16] Sijin Li and Antoni B. Chan. 3D human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 1, 2

[17] Sijin Li, Weichen Zhang, and Antoni Chan. Maximum-margin structured learning with deep networks for 3D human pose estimation. In *ICCV*, 2015. 1, 2

[18] Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. Learning the depths of moving people by watching frozen people. In *CVPR*, 2019. 2, 6

[19] Zhi Li, Xuan Wang, Fei Wang, and Peilin Jiang. On boosting single-frame 3d human pose estimation via monocular videos. In *ICCV*, 2019. 2, 8

[20] Diogo C Luvizon, David Picard, and Hedi Tabia. 2d/3d pose estimation and action recognition using multitask deep learning. In *CVPR*, 2018. 2

[21] Julieta Martinez, Rayat Hossain, Javier Romero, and James J. Little. A simple yet effective baseline for 3D human pose estimation. In *ICCV*, 2017. 2, 8

[22] 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 *3DV*, 2017. 1, 5, 6

[23] 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 *SIGGRAPH*, 2017. 8

[24] F. Moreno-Noguer. 3D human pose estimation from a single image via distance matrix regression. In *CVPR*, 2017. 2

[25] David Novotny, Nikhila Ravi, Benjamin Graham, Natalia Neverova, and Andrea Vedaldi. C3dpo: Canonical 3d pose networks for non-rigid structure from motion. In *ICCV*, 2019. 2

[26] Georgios Pavlakos, Nikos Kolotouros, and Kostas Daniilidis. Texturepose: Supervising human mesh estimation with texture consistency. In *ICCV*, 2019. 8

[27] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 1, 2, 3, 8

[28] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 1, 2

[29] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Harvesting multiple views for marker-less 3d human pose annotations. In *CVPR*, 2017. 1, 2, 8

[30] Dario Pavllo, Christoph Feichtenhofer, David Grangier, and Michael Auli. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *CVPR*, 2019. 2

[31] Gerard Pons-Moll, David J. Fleet, and Bodo Rosenhahn. Posebits for monocular human pose estimation. In *CVPR*, 2014. 2

[32] Alin-Ionut Popa, Mihai Zanfir, and Cristian Sminchisescu. Deep multitask architecture for integrated 2D and 3D human sensing. In *CVPR*, 2017. 1, 2

[33] Helge Rhodin, Mathieu Salzmann, , and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV*, 2018. 1, 2, 8

[34] 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 *CVPR*, 2018. 1, 2, 5, 6, 8

[35] Grégory Rogez and Cordelia Schmid. MoCap-guided data augmentation for 3D pose estimation in the wild. In *NeurIPS*, 2016. 2

[36] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. LCR-Net: Localization-classification-regression for human pose. In *CVPR*, 2017. 8

[37] Matteo Ruggero Ronchi, Oisin Mac Aodha, Robert Eng, and Pietro Perona. It's all relative: Monocular 3d human pose estimation from weakly supervised data. In *BMVC*, 2018. 2

[38] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[39] Xiao Sun, Jiaxiang Shang, Shuang Liang, and Yichen Wei. Compositional human pose regression. In *ICCV*, 2017. 1, 2, 3, 8

[40] Xiao Sun, Bin Xiao, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 1, 2, 3, 8

[41] Bugra Tekin, Isinsu Katircioglu, Mathieu Salzmann, Vincent Lepetit, and Pascal Fua. Structured prediction of 3D human pose with deep neural networks. In *BMVC*, 2016. 1, 2

[42] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3D pose estimation from a single image. In *CVPR*, 2017. 2, 8

[43] Hsiao-Yu Tung, Wei Tung, Ersin Yumer, and Katerina Fragkiadaki. Self-supervised learning of motion capture. In *NeurIPS*, 2017. 2

[44] Hsiao-Yu Fish 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 *ICCV*, 2017. 2

[45] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2

[46] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *CVPR*, 2019. 2, 8

[47] Chaoyang Wang, Chen Kong, and Simon Lucey. Distill knowledge from nrsfm for weakly supervised 3d pose learning. In *ICCV*, 2019. 2, 8

[48] Jiajun Wu, Tianfan Xue, Joseph J Lim, Yuandong Tian, Joshua B Tenenbaum, Antonio Torralba, and William T Freeman. Single image 3d interpreter network. In *ECCV*, 2016. 2

[49] B. Xiao, H. Wu, and Y. Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 6

[50] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy Ren, Hongsheng Li, and Xiaogang Wang. 3d human pose estimation in the wild by adversarial learning. In *CVPR*, 2018. 2, 8

[51] Yuan Yao, Yasamin Jafarian, and Hyun Soo Park. Monet: Multiview semi-supervised keypoint detection via epipolar divergence. In *ICCV*, 2019. 2

[52] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *ICCV*, 2017. 1, 2, 8

[53] Xingyi Zhou, Xiao Sun, Wei Zhang, Shuang Liang, and Yichen Wei. Deep kinematic pose regression. In *ECCV Workshops*, 2016. 1, 2