CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild

Bastian Wandt[1,2] Marco Rudolph[1] Petrissa Zell[1] Helge Rhodin[2]
Bodo Rosenhahn[1]
[1]Leibniz University Hannover, Hannover, Germany
[2]University of British Columbia, Vancouver, Canada
wandt@cs.ubc.ca

Abstract

*Human pose estimation from single images is a challenging problem in computer vision that requires large amounts of labeled training data to be solved accurately. Unfortunately, for many human activities (*e.g. *outdoor sports) such training data does not exist and is hard or even impossible to acquire with traditional motion capture systems. We propose a self-supervised approach that learns a single image 3D pose estimator from unlabeled multi-view data. To this end, we exploit multi-view consistency constraints to disentangle the observed 2D pose into the underlying 3D pose and camera rotation. In contrast to most existing methods, we do not require calibrated cameras and can therefore learn from moving cameras. Nevertheless, in the case of a static camera setup, we present an optional extension to include constant relative camera rotations over multiple views into our framework. Key to the success are new, unbiased reconstruction objectives that mix information across views and training samples. The proposed approach is evaluated on two benchmark datasets (Human3.6M and MPII-INF-3DHP) and on the in-the-wild SkiPose dataset.*

1. Introduction

Human pose estimation from single images is an ongoing research topic in computer vision. There exist a large amount of supervised deep learning solutions in the literature. These approaches achieve remarkable results in a supervised setting, *i.e.* having 2D to 3D annotations, but heavily rely on a vast amount of available training data. However, there are many activities a person can perform which are not present in common datasets. For instance, human motions performed during outdoor and/or sports activities, *e.g.* as shown in Fig. 1, are hard or even impossible to capture with a commercial motion capture systems. Therefore, the acquisition of training data is a major practical challenge. To this end, we propose a novel self-supervised training procedure that does not require any 2D or 3D annotations in the multi-view training dataset and works with uncalibrated cameras. To acquire 2D joint predictions from images we use a 2D human joint estimator [7] that is pretrained on a different dataset with only 2D joint annotations. The only requirements for our method are at least two temporally synchronised cameras that observe the person of interest from different angles. No further knowledge about the scene, camera calibration and intrinsics is required. Several related works consider a sparse set of 3D annotations [36, 34, 29], unpaired 3D data [47, 48, 16], or known camera positions [36, 34] to solve this problem. However, such data rarely exists for outdoor settings with moving cameras. To the best of our knowledge, there are only three competing methods [2, 14, 11] that apply to our setting. They either require additional knowledge about the scene or observed person, such as scene geometry [2] and bone lengths constraints [11], or sophisticated traditional computer vision algorithms that produce a pseudo ground truth pose [14].

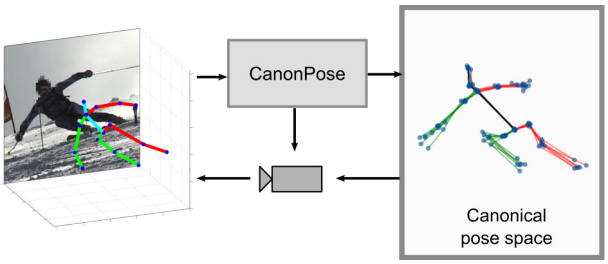

Figure 1. CanonPose learns a monocular 3D human pose estimator from multi-view self-supervision. By estimating 3D poses from different views in a canonical form together with the respective camera rotations we exploit multi-view consistency in the training data. Even for challenging outdoor datasets with moving cameras we achieve convincing 3D pose estimates from single images after training.

We propose a self-supervised training method which mixes outputs of multiple weight-sharing neural networks. Fig. 2 shows our training pipeline when using two cameras. Each individual network takes a single image as input and outputs a 3D pose in a canonical rotation, which gives our

method its name *CanonPose*. This representation allows for the projection of all estimated 3D poses to any camera of the setup. Our approach splits into two stages. The first stage predicts the 2D human pose from an image using a neural network pretrained on the MPII dataset [24], in our case AlphaPose [7, 17]. The second stage lifts these 2D detections to a 3D pose represented in a learned canonical coordinate system. In a separate path it predicts the camera orientation to rotate the predicted 3D pose back into the camera coordinate system. Combining the 3D pose from a first view with the rotation predicted from a second view, results in a rotated pose in the second camera coordinate system. In other words, both 3D poses in the pose coordinate system should be equal and the predicted rotations project it back into the respective camera coordinate systems. This enables the definition of a reprojection loss for each original and newly combined reprojection. For static camera setups we propose an optional reprojection loss that is computed by mixing relative camera rotations between samples in a training batch. Additionally, in contrast to existing self-supervised approaches, we make use of the confidences that are typically provided by 2D pose estimators for each predicted 2D joint by including them into the 2D input vector as well as into the reprojection loss formulation.

We evaluate our approach on the two benchmark datasets Human3.6M [10] and MPI-INF-3DHP [24] and set the new state of the art in several metrics for self-supervised 3D pose estimation. Notably, this is without assuming any camera calibration or static cameras. Our results are competitive to the fully supervised approach of Martinez et al. [23] which sets the baseline for single image pose estimation from 2D detections. Additionally, we show results for the SkiPose [39, 36] dataset. This dataset represents all challenges that arise when activities are captured that cannot be performed in the restricted setting of a standard motion capture system. It consists of outdoor scenes captured on a ski slope and includes fast motions, a large capture volume and pan-tilt-zoom cameras.

The code is available on GitHub [1].

Summarizing, our contributions are:

- We present CanonPose: a self-supervised approach to train a single image 3D pose estimator from unlabeled multi-view images by mixing poses across views.
- Our approach requires no prior knowledge about the scene, 3D skeleton or camera calibration.
- The proposed method directly employs multi-view images without any laborious pre-processing, such as camera calibration or multi-view geometry estimation.
- We integrate the confidences from the 2D joint estimator into the training pipeline.

[1] https://github.com/bastianwandt/CanonPose

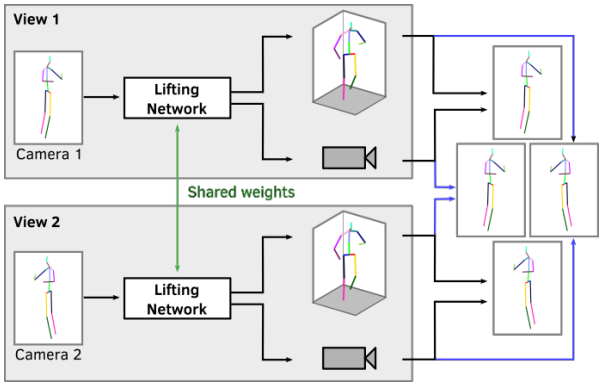

Figure 2. Network structure to learn single image 3D pose estimation from multi-view self-supervision. Each lifting network predicts a 3D pose and a camera rotation which is used to project the 3D pose back to 2D. Both networks observe the same 3D pose from different angles. We exploit this fact by applying the camera rotation to the respective other pose. This projects a predicted 3D pose into the other camera and gives an additional reprojection error. At inference time only one view (gray box) is applied.

2. Related Work

In this section we discuss recent 3D human pose estimation approaches by different types of supervision.

Full Supervision Recent supervised approaches rely on large datasets that contain millions of images with corresponding 3D pose annotations. Li et al. [18] were the first to learn CNNs to directly regress a 3D pose from image input. By integrating a structured learning framework into CNNs they later improved their work [20]. Several others followed this end-to-end approach [44, 30, 5, 26, 32, 38, 45, 42, 22, 43, 52, 19, 50, 13]. Typically, these end-to-end approaches perform exceptionally well on similar image data. However, their ability to generalize to other scenes is limited. Many works tackle this problem by cross dataset training or data augmentation.

There are other approaches that do not consider the image data directly but use a pretrained 2D joint detector [1, 28, 6, 24, 31, 25]. They benefit from training on large datasets that contain 2D annotations for many human activities in various scenes and are therefore agnostic to the image data. Martinez et al. [23] directly train a neural network on 2D detections and 3D ground truth. Due to its simplicity it can be trained quickly for many epochs leading to high accuracy and serves as a baseline for many following works. The approach of [23] was extended by Hossain et al. [33] by employing a recurrent neural network for sequences of human poses. While effective, the major downside of all supervised approaches is that they do not generalize well to unseen poses. Therefore, their application to in-the-wild scenes is limited.

Weak Supervision Weakly supervised approaches only require a small set or even no annotated 2D to 3D correspondences. An example for a commonly applied evaluation protocol for the Human3.6M dataset assumes that 3D annotations for one of the subjects of the training set are available. A transfer learning approach is introduced by Mehta et al. [24] to allow for in-the-wild pose estimation of datasets where no training data is available. This framework was later extended by Mehta et al. [26] to achieve real-time performance. Rhodin et al. [36] use multi-view images and known camera positions to learn a 3D pose embedding in an unsupervised fashion. The embedding facilitates the training with only a sparse set of 3D annotations. This idea was adopted in other works [34, 29, 27, 40]. Another approach is to employ unpaired 2D and 3D poses [47, 48, 16, 51, 3, 8]. Since these method learn distributions of plausible 3D poses and their properties they generalize better to unseen poses. Although they are able to reconstruct out-of-distribution poses to a limited degree they struggle with completely unseen poses.

Self-supervised and Unsupervised Learning without 3D Ground Truth Recently, the interest in multi-view self-supervised and unsupervised 3D pose estimation is growing. Our work also falls into this category. Drover et al. [4] propose an unsupervised 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. Chen et al. [2] extended [4] with a cycle consistency loss that is computed by lifting the randomly projected 2D pose to 3D and inversing the previously defined random projection. Although these two approaches are unsupervised, they integrate knowledge about the scene by constraining the camera rotation axis that is used for the random projection. Rochette *et al*. [37] use a large amount of cameras from different viewing angles to achieve on par performance with a comparable fully supervised approach. However, due to the restriction to the camera setup the practical applicability is limited. Kocabas et al. [14] propose a multi-view self-supervised approach which does not require any 3D supervision. They apply traditional computer vision methods, namely epipolar geometry, to 2D pose predictions from multiple views to compute a pseudo ground truth which is then used to train the 3D lifting network. Although this simple and effective straight-forward approach gives promising results, the laborious preprocessing step is very parameter sensitive and therefore does not generalize well. Moreover, mistakes due to wrongly estimated joints in the 2D prediction step result in a wrong pseudo ground truth. Iqbal et al. [11] tackle this problem by training an end-to-end network that refines the pre-trained 2D pose estimator during the self-supervised training. Unfortunately, such approaches tend to easily overfit to a specific dataset. For example, it could learn a background image for the training dataset which leads to exceptional performance on the specific dataset but does not generalize to other backgrounds. This even happens in self-supervised settings. Furthermore, Iqbal et al. [11] employ a loss on normalized 3D bone lengths which is computed from the ground truth 3D poses of the Human3.6M training set.

In contrast, our approach does not require knowledge about the scene and camera position or any anthropometric constraints. Additionally, it is modular such that any 2D pose estimator can be used which makes it agnostic to the image data. Even though our approach relaxes many constraints of the comparable works it still outperforms them in most experiments.

3. Method

Our approach consists of two steps: first applying an off-the-shelf 2D joint detector to the input images, and second lifting these detections and the respective confidences for each joint to 3D. The core idea of our approach is that 2D detections from one view can be projected to another view via a canonical pose space[2]. Fig. 2 shows our pipeline using two cameras. For simplicity the network structure is shown for only two cameras. If more cameras are available it is straight-forwardly extended. A single neural network, the *3D lifting network*, predicts the 3D pose $\boldsymbol{X} \in \mathbb{R}^{3\times j}$ with j joints and a rotation $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ to rotate the pose to the camera coordinate system. The pose is represented in a canonical pose coordinate system which is automatically learned during training. Subsequently, the predicted 3D pose is rotated from the pose coordinate system to the camera coordinate system by the predicted rotation. This separation into canonical human pose and camera rotation enables us to formulate various reprojection losses for self-supervision across views and samples.

3.1. Reprojection

Before a 2D pose is lifted to 3D it is normalized by centering it to the root joint and scaled by dividing it by its Frobenius norm. This sidesteps the scale-depth ambiguity in monocular reconstruction. In particular, the root centering gives a common rotation point for all 3D predictions. For each view the predicted 3D pose is rotated into the camera coordinate system by $\boldsymbol{RX}$. $\boldsymbol{R} \in \mathbb{R}^{3\times 3}$ is a rotational matrix such that $\boldsymbol{R}\boldsymbol{R}^T = \boldsymbol{I}_3$ with $\boldsymbol{I}_3$ as the 3×3 identity matrix and $det(\boldsymbol{R}) = 1$. Since we assume weak perspective cameras, the projection to the camera plane is simply done by removing the depth coordinate, which is expressed as

$$\boldsymbol{W}_{\text{rep}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{RX}, \tag{1}$$

[2]The benefit of canonical representations is also shown in [41].

where $\boldsymbol{W}_{rep} \in \mathbb{R}^{2\times j}$ is called the reprojected 2D pose. With $\boldsymbol{W}$ as the input 2D pose we define the *scale-independent reprojection loss* as

$$\mathcal{L}_{\text{rep}} = \left\| \boldsymbol{W} - \frac{\boldsymbol{W}_{\text{rep}}}{\|\boldsymbol{W}_{\text{rep}}\|_F} \right\|_1, \tag{2}$$

where $\|\cdot\|_1$ denotes the L_1 norm. Since the global scale of the 3D pose is unknown and we consider weak perspective projections, scaling the reprojection $\boldsymbol{W}_{\text{rep}}$ is essential. Note that the input 2D pose $\boldsymbol{W}$ is already divided by its Frobenius norm in the preprocessing. That means both, the input pose and the predicted pose, have the same scale.

To ensure that the network predicts a proper rotation, the matrix $\boldsymbol{R}$ is not predicted directly, but in axis-angle representation. Let (θ) be a rotational angle and $\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3)$ denote a rotation axis. With

$$\boldsymbol{A} = \begin{pmatrix} 0 & -\omega_3 & \omega_2 \\ \omega_3 & 0 & -\omega_1 \\ -\omega_2 & \omega_1 & 0 \end{pmatrix} \tag{3}$$

Rodrigues' formula is applied to obtain the rotation matrix

$$\boldsymbol{R} = \boldsymbol{I}_3 + (\sin\theta)\boldsymbol{A} + (1-\cos\theta)\boldsymbol{A}^2. \tag{4}$$

3.2. View-consistency

A straight-forward way of ensuring view consistency would be to enforce a loss, such as L_2, between the canonical poses predicted by two views. In theory, that loss should be zero for the correct solution because the same person seen from two different views should have the same canonical pose. In practice, however, this leads to the lifting network learning 3D poses that are view invariant but no longer in close correspondence to the input pose, preventing the network to converge to plausible solutions in our preliminary experiments.

The key insight to the proposed method is that rotations and poses from different views can be mixed to enforce the view consistency as a variant of the previously introduced reprojection objective. We mix the predicted camera and pose of two views, say view-1 and view-2, by rotating the predicted canonical 3D pose from the source view-1 to the target view-2 by using the rotation from view-2. For two cameras as in Fig. 2 there exist four possible combinations of rotations and poses. The same approach is easily extended to m cameras which gives m^2 combinations. During training time all possible combinations are reprojected to the respective cameras. With this training scheme we enforce multi-view consistency without bias towards trivial solutions. Note that the lifting network is only applied to a single frame at inference stage and does not need any other inputs.

3.3. Confidences

The output of most pretrained 2D joint estimators are 2D heatmaps where each entry indicates the confidence for the presence of the corresponding joint at the associated position in the image. Commonly, the argmax or soft-argmax is computed and given as input to the following lifting network. However, this gives an exact joint position independent of the confidence of the 2D detection. That means uncertain predictions are processed in the same way as certain ones. We circumvent this problem by two simple modifications. First, we concatenate the maximum value of each heatmap, which is a surrogate to its confidence, to the 2D pose input vector to our lifting network. Second, we modify the reprojection error in Eq. 2 such that each difference between input and reprojected 2D is linearly weighted with its confidence by

$$\mathcal{L}_{\text{rep,c}} = \left\| \left(\boldsymbol{W} - \frac{\boldsymbol{W}_{\text{rep}}}{\|\boldsymbol{W}_{\text{rep}}\|_F} \right) \odot \boldsymbol{C} \right\|_1, \tag{5}$$

where

$$\boldsymbol{C} = \begin{pmatrix} c_1 & c_2 & \dots & c_j \\ c_1 & c_2 & \dots & c_j \end{pmatrix} \tag{6}$$

with c_i as the maximum value of the heatmap for joint i and $\odot$ as the Hadamard product.

3.4. Camera-consistency

A reasonable assumption for many practical motion capture setups is that cameras are static during recording a sequence, *i.e.* they do not change their position or orientation. This is the case for the Human3.6M[3] and 3DHP dataset. However, this assumption is not mandatory for our proposed method, but an enhancement for scenes with static cameras. We will show the effect of this optional improvement in the experiments as well as the performance of our approach without it on the SkiPose dataset that contains moving cameras.

For a static camera setup all relative rotations between the cameras are equal. An intuitive approach to enforce static cameras is to calculate an L_2-loss between the relative rotations over one batch of training samples. However, a batch-wise loss leads to degraded solutions or had no effect if its weight was set to a low value. This observation is similar to the findings regarding the canonical pose equality in Sec. 3.2. For this reason we propose a similar mixing approach as in Sec. 3.2, now over estimates from different samples in one batch. A relative rotation $\boldsymbol{R}_{1,2}$ using the rotation matrices $\boldsymbol{R}_1$ and $\boldsymbol{R}_2$ from view-1 to view-2 respectively, is defined by

$$\boldsymbol{R}_{1,2} = \boldsymbol{R}_2 \boldsymbol{R}_1^T. \tag{7}$$

[3] In fact, camera angles change between subjects but not during a capture session with one subject.

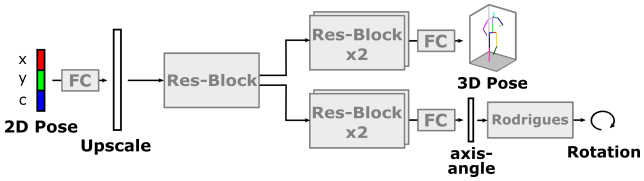

Figure 3. Network structure of the lifting network. The 2D input vector contains the x- and y-coordinates of the 2D pose and the confidence given by the 2D joint detector. It is upscaled using a fully connected layer with 1024 neurons which then goes to a residual block. After that the network splits into two paths that predict the 3D pose in the canonical space and the camera rotation, respectively. Each of the paths has two consecutive residual blocks followed by a fully connected layer that downscales the features to the required size. The Rodrigues block implements Rodrigues formula (Eq. 4) and has no trainable parameters.

Let $\boldsymbol{R}_{1,2}^{(s)}$ be the predicted relative rotation between view-1 and view-2 of sample s. We then randomly permute these relative rotations in the batch and use them to reproject the canonical poses similar to Eq. 1

$$\boldsymbol{W}_{\text{rep}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \boldsymbol{R}_{1,2}^{(s)} \boldsymbol{R}_{1}^{(s')} \boldsymbol{X}^{(s')}, \tag{8}$$

where $\boldsymbol{R}_{1}^{(s')}$ and $\boldsymbol{X}^{(s')}$ are the rotation and estimated 3D pose in the current frame and $\boldsymbol{R}_{1,2}^{(s)}$ is the randomly assigned relative rotation from another sample in the batch[4]. The loss is calculated in the same way as the reprojection loss in Eq. 2. Similar to Sec. 3.2 this is easily extended to multiple cameras. Again, we emphasize that this loss is optional to improve the results for the case of static cameras. However, our method works without it.

3.5. Network Architecture

Fig. 3 shows the architecture of our lifting network. The input 2D pose vector is concatenated with a vector containing the confidences for each joint. It is upscaled to 1024 neurons by a one fully connected layer. It is followed by a residual block consisting of fully connected layers with dimension 1024. Similar to [47] the output is fed into two paths, each containing two consecutive residual blocks with identical architecture to the first block. The 3D pose path directly outputs the 3D coordinates of the predicted pose in the pose coordinate system. The camera path outputs a three-dimensional vector $\theta\boldsymbol{\omega}$ which is the axis angle representation. The rotation matrix is computed using Rodrigues' formula as described in Sec. 3.1. The activation functions after each layer, except the two output layers, are leaky ReLU's with a negative slope of 0.01. We train the network for 100 epochs using the Adam optimizer with an initial learning rate of 0.0001 and weight decay at epochs 30, 60 and 90, respectively.

[4]For the Human3.6M dataset we ensure that relative rotations are only changed in between subjects since camera positions vary between them.

4. Experiments

We perform experiments on the well-known benchmark datasets Human3.6M [10] and MPI-INF-3DHP [24]. Additionally, we evaluate on the SkiPose dataset [39, 36] to test the generalizability of our method to real world scenarios. To conform with our setting of training a single image pose estimator with unlabeled images for a specific set of activities, we train one network for each dataset without using additional datasets.

4.1. Metrics

For the evaluation on Human3.6M there exist two standard protocols. Both protocols calculate the *mean per joint position error* (MPJPE), i.e. the mean euclidean distance between the reconstructed and the ground truth joint coordinates. Since a multi-view self-supervised setting does not contain metric data, we adjust the scale of our predictions before calculating the MPJPE. For a fair comparison with other works we compare to their scale adjusted predictions if they are available. Protocol-I computes the MPJPE directly whereas Protocol-II first employs a rigid alignment between the poses. Additional to the MPJPE one protocol for 3DHP calculates the *Percentage of Correct Keypoints* (PCK). As the name suggests it is the percentage of predicted joints that are within a distance of $150mm$ or lower to their corresponding ground truth joint.

Correct Poses Score (CPS)

For practical applications, such as motion analysis and prediction, the evaluation of the whole pose is a crucial prerequisite. Even if a single joint of a pose is incorrect it can change downstream tasks significantly. The formerly introduced metrics evaluate the quality of the prediction joint by joint. However, they ignore the assignment of joints to poses and instead average over all joints in the test set. Fig. 5 compares 3D pose estimates with their respective ground truths. Each column shows two different reconstructions from the same pose. The reconstructions in the top row have a lower PMPJPE compared to the bottom row. However, the overall 3D poses appear better reconstructed in the bottom row. In this section we present a simple yet powerful metric to evaluate such cases, the *Correct Poses Score* (CPS). A pose $\boldsymbol{W}$ is considered correct if for all joints i the Euclidean distance is below a threshold value θ. Given a pose with joint positions $\boldsymbol{w}_i$ and predicted joint positions $\hat{\boldsymbol{w}}_i$ after rigid alignment, a correct pose is defined by

$$\text{CP}_\theta = \begin{cases} 1 & \|\boldsymbol{w}_i - \hat{\boldsymbol{w}}_i\|_2 < \theta \quad \forall i \in \{1, ..., j\} \\ 0 & else \end{cases}. \tag{9}$$

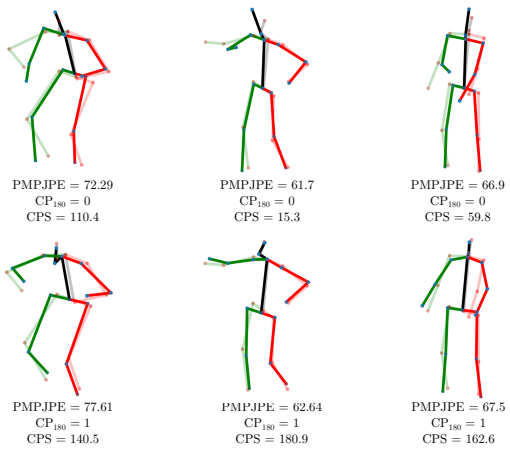

Figure 4. Comparison of PMPJPE and our CP-metric. Each column compares to different predicted 3D reconstructions with the same ground truth. While PMPJPE averages out high individual joint errors which are located in the right arm in the visualized case, CP indicates them. In this way, the correctness of the overall pose is evaluated. Note that for the calculation of the CPS we vary the threshold, which in these examples is $180mm$.

Additional to the PMPJPE Fig. 5 shows the CP@180mm, which classifies the reconstructed poses into correct and incorrect. The percentage of correct poses is calculated for the test dataset. To be independent of the threshold, we calculate the area under curve for $\theta \in [0mm, 300mm]$ which defines the CPS.

4.2. Skeleton Morphing

We deploy an off-the-shelf detector AlphaPose [7] for retrieving the 2D human pose estimation required as input to our method. The keypoint locations in the datasets used to train AlphaPose and other 2D pose estimation methods differ from the 3D skeleton of the test benchmarks. For example, the root joint position is not in the middle of the hip joints and the relative position of the neck to the shoulders is different. We circumvent this problem by training a *2D skeleton morphing network* that predicts the offset between the 2D pose from AlphaPose to the ground truth 2D pose in the dataset. We train the morphing network on subject 1 of each dataset with the given ground truth poses. To not include these ground truth poses into our training, subject 1 is excluded in all experiments. Thereby our data used for the self supervised training does not contain any 2D ground truth data, mimicking real application scenarios. Note that the morphing network never sees any images and therefore is not able to learn domain specific image features. In an experimental setting where the skeletal structure does not need to match a different skeleton this step is obsolete. This is the case for most practical applications.

Table 1. Evaluation results for the Human3.6M dataset in mm. The bottom section, labeled with *self*, shows methods that can solve our setting. Best results are marked in bold and second best in italic.

Supervision	Method	MPJPE↓	PMPJPE↓
full	Martinez [23]	67.5	52.5
weak	Rhodin [36]	80.1	65.1
	Rhodin [35]	122.6	98.2
	3D interpreter [49]	98.4	88.6
	AIGN [46]	97.2	79.0
	RepNet [47]	89.9	65.1
	HMR [12]	-	66.5
	Wang [48]	86.4	62.8
	Kolotouros [15]	-	62.0
	Kundu [16]	85.8	-
self	Chen [2]	-	68.0
	EpipolarPose [14]	76.6	67.5
	Iqbal [11]	**69.1**	55.9
	Ours	81.9	**53.0**
	Ours + C	*74.3*	**53.0**

4.3. Quantitative Evaluation on Human3.6M and 3DHP

For the Human3.6M dataset, to keep it consistent with previous approaches, we follow standard protocols and evaluate only on every 64th frame. However, with a sufficiently fast 2D pose estimator, which is the performance bottle neck of our complete pipeline, we can achieve real-time performance. Table 1 shows the results of the proposed method compared to other state-of-the-art approaches. We outperform every other comparable approach in terms of PMPJPE. Note that we even achieve comparable performance to the fully supervised method of Martinez *et al.* [23] which has a lifting network with similar structure to ours. Only one other self-supervised approach attains a lower MPJPE, however, by using additional information. Our analysis revealed that although our pose structure is very accurate (which results in a low PMPJPE) the largest part of the error originates from a slight offset in the rotation. For example, comparing frame 1 from subject 9 of the Human3.6M dataset to itself rotated by only $15°$ around the longitudinal axis already results in an MPJPE of $67.7mm$. Most methods [2, 14, 11] that have no knowledge of the scene or any 3D training data show this large gap between MPJPE and PMPJPE, as a small rotation leads to small reprojection loss but large 3D MPJPE error. Iqbal *et al.* [11] still set the state of the art in terms of MPJPE. However, they need bone length constraints which they directly compute from the ground truth 3D data of the training set. Our approach does not require any predefined priors on the skeletal structure. Using our static camera constraint (Ours+C) improves the MPJPE significantly.

Fig. 5 shows the CPS for our method compared to EpipolarPose [14], which is the only comparable approach with publicly available code, and the 3D pose estimation baseline of Martinez et al. [23]. On this metric, we outper-

Table 2. Evaluation results for the 3DHP dataset. The bottom section, labeled with *self*, shows methods that can solve our setting. Best results are marked in bold and second best in italic. MPJPE and PMPJPE are given in mm, PCK is in $\%$.

Supervision	Method	MPJPE↓	PMPJPE↓	PCK↑
weak	Rhodin [36]	121.8	-	72.7
	HMR [12]	169.5	-	59.6
	Habibie [9]	-	-	70.4
	Kolotouros [15]	124.8	-	66.8
	Li [21]	-	-	74.1
	Kundu [16]	103.8	-	82.1
self	Chen [2]		71.1	
	EpipolarPose [14]	125.7	-	64.7
	Iqbal [11]	*110.1*	-	*76.5*
	Ours	119.2	**68.7**	69.0
	Ours + C	**104.0**	*70.3*	**77.0**

form EpipolarPose by a large margin. Note the high threshold of over $80mm$ that is required by [14] to achieve a CP above 1% compared to our threshold slightly below $50mm$. As for the CPS metric, we are on par with the fully supervised approach of [23]. Since their originally trained model is not publicly available anymore we retrained their model with their provided code to report the new CPS metric. The retrained model achieved a PMPJPE of 53.5mm, which is slightly lower compared to their original number. The new model is used only for reporting CPS. Fig. 6 shows qualitative results for the Human3.6M data set in the first row.

We also evaluate our approach on the 3DHP dataset [24] following the standard test protocols and metrics. Table 2 shows the results. We outperform every other self-supervised approach. In contrast to other approaches the proposed method does not require calibrated cameras[5] or anthropometric constraints. For the CPS metric we achieve a score of 134.2.

4.4. Moving cameras

Our main motivation is to enable 3D human pose estimation in the wild by using a multi-view camera system with temporally synchronised cameras. Moreover, the performed activity should be very challenging to capture and hard to simulate in a traditional motion capture studio. That means a straight-forward activity domain transfer, *e.g.* pretraining or combined training with a different dataset, is not reasonable. The SkiPose dataset [39, 36] comprises all challenges of this motivation. It features competitive alpine skiers performing giant slalom runs. To record this dataset huge effort was taken to setup and calibrate the cameras and keep them in place after calibration. Additionally, the cameras are rotating and zooming to keep the alpine skier in the field of view. The proposed method can deal with all these difficulties since it does not require a calibrated or static setup and works with multiple synchronised cameras. Since the camera setup is not static we cannot apply the relative rotation constraint here. Table 3 shows our results for different configurations in comparison to Rhodin *et al.* [36]. Since they consider a (sparse-)supervised setting and known camera positions a direct comparison is not possible and only serves as a baseline. Fig. 6 shows qualitative results for the SkiPose dataset in the second row.

[5]The configuration Ours+C only assumes that cameras are static during the sequence, which is a much weaker constraint.

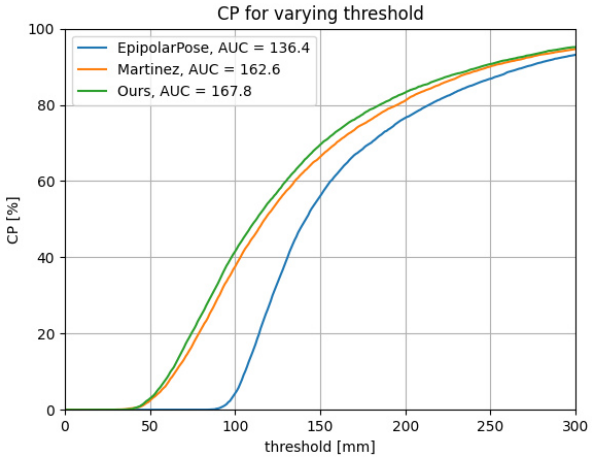

Figure 5. Comparison of CPS curves for distances from $1mm$ to $300mm$ with corresponding AUC for the Human3.6M dataset. A higher value means a better result, *i.e.* the leftmost curve achieves the best result in terms of CP.

Table 3. Evaluation results for the SkiPose dataset. The result for [36] was estimated from a bar plot in the paper. Since [36] considers a (sparse-)supervised setting and known camera position it is only shown as a baseline. MPJPE, PMPJPE and CPS are given in mm, PCK is in $\%$.

Supervision	Method	MPJPE↓	PMPJPE↓	PCK↑	CPS↑
weak	Rhodin [36]	85	-	-	-
self	2 cams (Ours)	201.9	122.4	47.4	54.8
	3 cams (Ours)	176.9	106.7	54.5	82.8
	4 cams (Ours)	139.3	95.8	61.9	94.3
	5 cams (Ours)	129.9	90.7	66.4	106.9
	w/o conf. (Ours)	135.7	95.5	58.8	79.3
	full (Ours)	128.1	89.6	67.1	108.7

4.5. Ablation Studies

To analyze our approach we perform a number of ablation studies. First, to simulate a practical setting with limited resources, we reduced the number of cameras to train the model. Table. 4 and Table 3 show results for the training with different numbers of cameras. While the performance expectedly slightly drops due to the lower number of training samples and views our approach still produces good results which underlines its applicability in real world scenarios. In a second experiment we show the impact of using the confidences from the 2D joint estimator as inputs to the network and for the calculation of the reprojection error. They significantly improve the performance of our model.

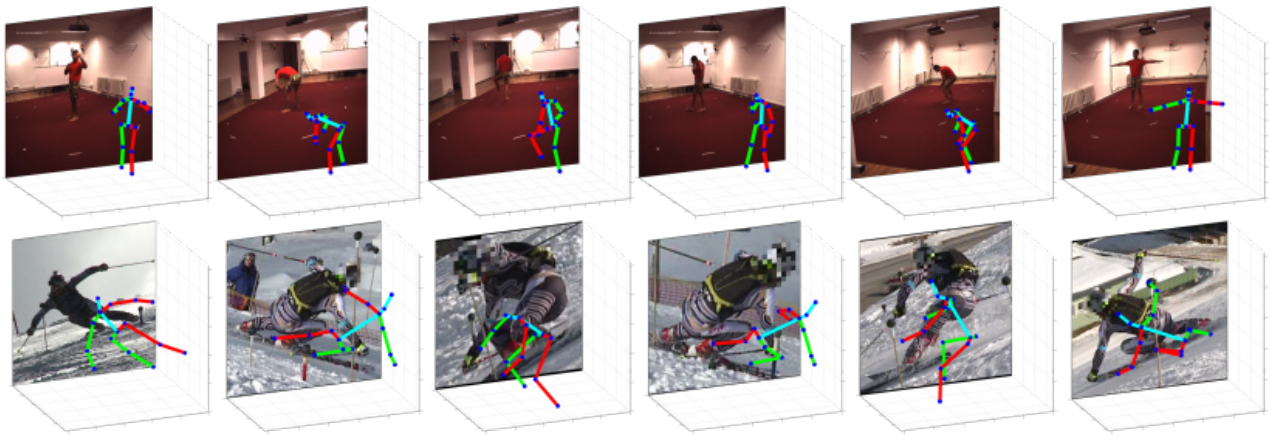

Figure 6. Qualitative results for the Human3.6M dataset (top) and for the challenging SkiPose dataset (bottom).

Table 4. Ablation studies on the Human3.6M dataset. All values are given in mm.

	MPJPE↓	PMPJPE↓	CPS↑
2 cams	82.7	61.2	148.5
3 cams	82.0	62.2	145.6
w/o confidences	95.6	65.0	142.5
ground truth 2D	65.9	51.4	187.1
direct pose equality	554.3	360.8	0.0
direct camera equality	617.9	374.5	0.0
full (4 cams)	81.9	53.0	167.6
full+C (4 cams)	74.3	53.0	167.3

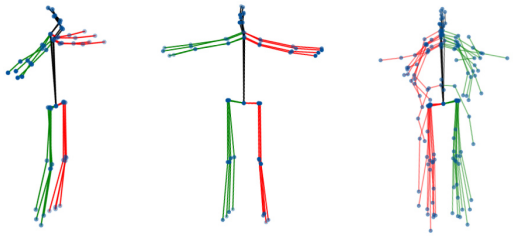

Figure 7. Visualization of the canonical pose space from the Human3.6M dataset. Left and middle: Canonical poses for the same 3D pose predicted from 4 different views. Right: 10 randomly sampled canonical poses. Our network automatically learns a disentanglement of a 2D pose into 3D and a camera rotation.

To prove that the proposed mixing of rotations and poses to achieve view- and camera-consistency is superior to simple equality constraints, we performed experiments with such equality constraints. The results show that indeed our mixing approach is an essential part to make it work. We also experimented with the bone lengths constraints from [11] that, however, did not improve the results. To compute a lower bound we also show results for training with ground truth 2D annotations.

4.6. Are We Learning a Canonical Pose Basis?

Finally, we evaluate the claim that we learn a canonical pose basis. To visualize the disentanglement for different 3D poses Fig. 7 shows a visualization of reconstructed 3D poses in the canonical basis obtained from 4 views on the left and in the middle. The right image shows 10 randomly picked reconstructions in the canonical space. Although the similarity of the poses is not enforced directly as described in Sec. 3.2 the poses are similarly oriented in the canonical space. In particular, the hip joints are aligned which leads to a similar alignment of the upper body. The standard deviation for the hip joints of the canonical poses from the test set of Human3.6M are $7.9mm$ and $7.7mm$ for the right and left hip, respectively. This underlines that pose and rotation are disentangled plausibly by our network.

5. Conclusion

We present CanonPose, a neural network trained for single image 3D human pose estimation from multi-view data without 2D or 3D annotations. Given a pretrained 2D human pose estimator we exploit multi-view consistency to automatically decompose a 2D observation into a canonical 3D pose and a camera rotation that is used to reproject it back to the observation after mixing. Since our approach does not require either 2D nor 3D annotations for the multi-view data it is practically applicable to many in-the-wild scenarios, including outdoor scenes with moving cameras. We not only achieve state-of-the-art results on benchmark datasets with less prerequisites compared to other approaches, but also show promising results on challenging outdoor scenes.

6. Acknowledgements

This work was partially supported by the Federal Ministry of Education and Research (BMBF), Germany under the project LeibnizKILabor (grant no. 01DD20003) and by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy within the Cluster of Excellence PhoenixD (EXC 2122).

References

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

[2] 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, 6, 7

[3] 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. 3

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

[5] 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

[6] 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

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. RMPE: Regional multi-person pose estimation. In *ICCV*, 2017. 1, 2, 6

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

[9] 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. 7

[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. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(7):1325–1339, 2014. 2, 5

[11] 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. 1, 3, 6, 7, 8

[12] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7122–7131. IEEE Computer Society, 2018. 6, 7

[13] 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

[14] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 3, 6, 7

[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 *Proceedings International Conference on Computer Vision (ICCV)*, pages 2252–2261. IEEE, Oct. 2019. ISSN: 2380-7504. 6, 7

[16] Jogendra Nath Kundu, Siddharth Seth, Varun Jampani, Mugalodi Rakesh, R. Venkatesh Babu, and Anirban Chakraborty. Self-supervised 3d human pose estimation via part guided novel image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3, 6, 7

[17] Jiefeng Li, Can Wang, Hao Zhu, Yihuan Mao, Hao-Shu Fang, and Cewu Lu. Crowdpose: Efficient crowded scenes pose estimation and a new benchmark. *arXiv preprint arXiv:1812.00324*, 2018. 2

[18] 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

[19] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[20] 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

[21] Yang Li, Kan Li, Shuai Jiang, Ziyue Zhang, Congzhentao Huang, and Richard Yi Da Xu. Geometry-driven self-supervised method for 3d human pose estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):11442–11449, Apr. 2020. 7

[22] 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

[23] 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, 6, 7

[24] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *International Conference on 3D Vision (3DV)*. IEEE, 2017. 2, 3, 5, 7

[25] 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

[26] 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, 3

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

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

[29] 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. 1, 3

[30] 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

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

[32] 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

[33] 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

[34] Helge Rhodin, Victor Constantin, Isinsu Katircioglu, Mathieu Salzmann, and Pascal Fua. Neural scene decomposition for multi-person motion capture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7703–7713, 2019. 1, 3

[35] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation learning for 3d human pose estimation. In *ECCV*, 2018. 6

[36] 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. 1, 2, 3, 5, 6, 7

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

[38] 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

[39] Jörg Spörri. Reasearch dedicated to sports injury prevention-the'sequence of prevention'on the example of alpine ski racing. *Habilitation with Venia Docendi in Biomechanics*, 1(2):7, 2016. 2, 5, 7

[40] 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. 3

[41] Weiwei Sun, Andrea Tagliasacchi, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geoffrey Hinton, and Kwang Moo Yi. Canonical capsules: Unsupervised capsules in canonical pose, 2020. 3

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

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

[44] 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

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

[46] 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

[47] 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. 1, 3, 5, 6

[48] 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. 1, 3, 6

[49] 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

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

[51] Andrei Zanfir, Eduard Gabriel Bazavan, Hongyi Xu, William T. Freeman, Rahul Sukthankar, and Cristian Smin-

chisescu. 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. 3

[52] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2344–2353, 2019. 2