In IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) 2019

Unsupervised 3D Pose Estimation with Geometric Self-Supervision

Ching-Hang Chen[1] Ambrish Tyagi[1] Amit Agrawal[1] Dylan Drover[1]
Rohith MV[1] Stefan Stojanov[1,2] James M. Rehg[1,2]

[1]Amazon Lab126, [2]Georgia Institute of Technology

{chinghc, ambrisht, aaagrawa, droverd, kurohith}@amazon.com {sstojanov, rehg}@gatech.edu

Abstract

We present an unsupervised learning approach to recover 3D human pose from 2D skeletal joints extracted from a single image. Our method does not require any multi-view image data, 3D skeletons, correspondences between 2D-3D points, or use previously learned 3D priors during training. A lifting network accepts 2D landmarks as inputs and generates a corresponding 3D skeleton estimate. During training, the recovered 3D skeleton is reprojected on random camera viewpoints to generate new 'synthetic' 2D poses. By lifting the synthetic 2D poses back to 3D and re-projecting them in the original camera view, we can define self-consistency loss both in 3D and in 2D. The training can thus be self supervised by exploiting the geometric self-consistency of the lift-reproject-lift process. We show that self-consistency alone is not sufficient to generate realistic skeletons, however adding a 2D pose discriminator enables the lifter to output valid 3D poses. Additionally, to learn from 2D poses 'in the wild', we train an unsupervised 2D domain adapter network to allow for an expansion of 2D data. This improves results and demonstrates the usefulness of 2D pose data for unsupervised 3D lifting. Results on Human3.6M dataset for 3D human pose estimation demonstrate that our approach improves upon the previous unsupervised methods by 30% and outperforms many weakly supervised approaches that explicitly use 3D data.

1. Introduction

Estimation of 3D human pose from images and videos is a classical ill-posed inverse problem in computer vision with numerous applications [12, 17, 28, 31] in human tracking, action understanding, human-robot interaction, augmented reality, video gaming, *etc*. Current deep learning-based systems attempt to learn a mapping from RGB images or 2D keypoints to 3D skeleton joints via some form of supervision requiring datasets with *known 3D pose*. However, obtaining 3D motion capture data is time-consuming, difficult, and expensive, and as a result, only a limited amount of 3D data is currently available. On the other hand, 2D image and video data of humans is available in abundance. However, unsupervised learning of 3D joint locations from 2D pose alone remains a holy grail in the field. In this paper, we take a first step towards achieving this goal and present an unsupervised learning algorithm to estimate 3D human pose from 2D pose landmarks/keypoints. Our approach does not use 3D inputs in *any* form and does not require 2D-3D correspondences or explicit 3D priors.

Due to perspective projection ambiguity, there exists an infinite number of 3D skeletons corresponding to a given 2D pose. However, all of these solutions are not physically plausible given the anthropomorphic constraints and joint angle limits of a human body articulation. Typically, supervised learning with 2D pose and corresponding 3D skeletons is used to restrict the solution space. In addition, the 3D structure can also be regularized in a weakly-supervised manner by using priors such as symmetry, ratio of length of various skeleton elements, and kinematic constraints, which are learned from 3D data. In contrast, this paper addresses the fundamental problem of lifting 2D image coordinates to 3D space without the use of any additional cues such as video [43, 54], multi-view cameras [1, 16], or depth images [35, 40, 52].

We posit that the following properties of the 2D-3D pose mapping render unsupervised lifting possible: 1) *Closure:* If a 2D skeleton is lifted to 3D accurately, and then randomly rotated and reprojected, the resulting 2D skeleton will lie within the distribution of valid 2D poses. Conversely, a lifted 3D skeleton whose random re-projection falls outside this distribution is likely to be inaccurate. 2) *Invariance:* 2D projections of the same 3D skeleton from different viewpoints, when lifted, should produce the same 3D output. In other words, lifting should be invariant to change in the viewpoint.

We employ the above properties in designing a deep neural network, referred to as the *lifting network*, which is illustrated in Figure 1. We introduce a novel geometrical consistency loss term that allows the network to learn in a self-supervised mode. This self-consistency loss relies on the property of invariance: any 2D projection of the gen-

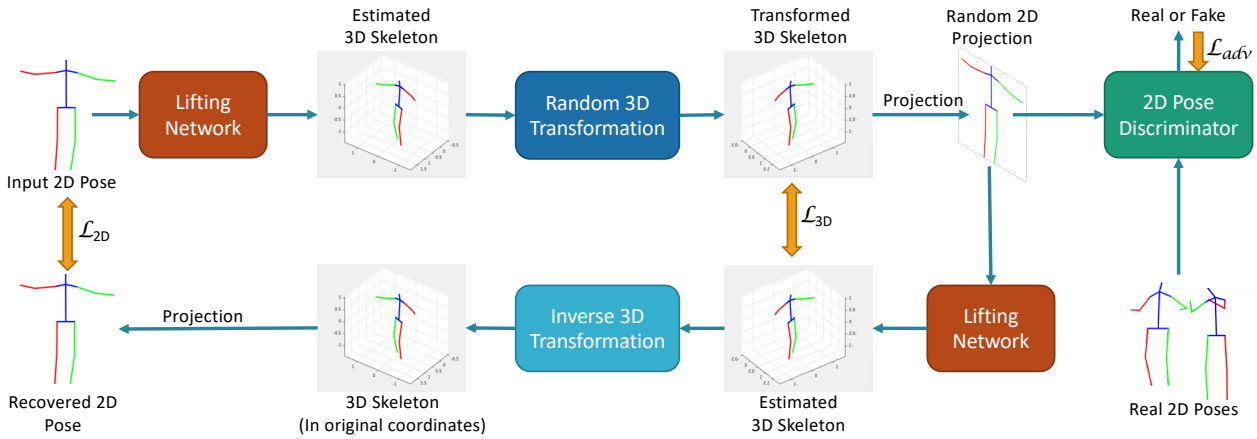

Figure 1. We train a 2D-3D lifting network (lifter), which estimates the 3D skeleton from 2D pose landmarks. Random projections of generated 3D skeletons are fed to a 2D pose discriminator to provide feedback to the lifter. The random projections also go through a similar lifting and reprojection process, allowing the network to self supervise the training process by exploiting geometric consistency.

erated 3D skeleton should produce the same 3D skeleton when processed by the lifting network (Section 3.3). We further demonstrate that self-consistency is a necessary but not a sufficient condition. We add a discriminator to ensure that the projection of lifted skeletons lie within the distribution of 2D poses. However we find that self-supervision *does* improve performance of the lifting network when used in conjunction with discriminator feedback.

Domain Adaptation: Since unsupervised learning methods often need more data for training than supervised methods, it is desirable to exploit multiple data sources. However, *domain shifts* could occur in multiple data sources due to (a) differences in human pose and viewpoint variations, and (b) semantic differences in the location of the skeletal joints on the body (e.g., hips marked inside/outside legs). We propose a domain adaptation algorithm, where we train a 2D *adapter* network to convert 2D joints from a source domain to the target domain without the need for any correspondences. Using multiple datasets allows us to enrich the viewpoint, pose, and articulation variations in the target domain using additional domains.

Temporal consistency during training: Our algorithm only requires 2D joints extracted from a single frame for training and inference. However, if sequences of poses from videos are available during training, we show how they can be used to improve the lifter. To exploit temporal consistency during training, we incorporate an additional *temporal discriminator* that classifies the differences in 2D joints in subsequent frames as real/fake. Our ablation studies show that adding a temporal discriminator improves performance by an additional 7%, even when inference is performed on a single frame.

Our paper makes the following contributions:

- Inspired by [9], we present an unsupervised algorithm to lift 2D joints to 3D skeletons by observing samples of real 2D poses, without using 3D data in any form.
- Our method can learn by exploiting geometric self consistency. We show that self consistency is a necessary but not a sufficient condition for lifting. Self consistency loss improves performance when combined with 2D pose discriminator adversarial loss.
- We propose a 2D domain adaptation technique which can utilize data from different domains to improve performance on the target domain.
- We show that adding a temporal discriminator during training can further improve performance, even for single frame 2D-3D lifting during inference.

2. Related Work

3D Pose Estimation: There are numerous deep learning techniques proposed for estimating 3D joint location directly from 2D images [7, 27, 32, 33, 34, 38]. Other methods decompose this problem into the estimation of 2D joint locations from images followed by the estimation of 3D joint locations based on the 2D keypoints. 2D pose from images can be obtained using techniques such as CPM [47], Stacked-hourglass architecture [30], Mask-RCNN [14] or affinity models [4]. As discussed, our focus is on estimating 3D pose from 2D landmarks [5, 11, 26], and we are agnostic to the source of landmarks. For the purpose of comparison, prior work on lifting can be organized into four categories:

Fully Supervised: These include approaches such as [25, 26, 49] that use paired 2D-3D data comprised of ground truth 2D locations of joint landmarks and the corresponding 3D ground truth for learning. For example, Martinez *et al*. [26] learn a regression network from 2D joints to 3D joints, whereas Moreno-Noguer [29] learn a regression from 2D distance matrix to 3D distance matrix using 2D-3D correspondences. Exemplar based methods [5, 19, 51] use a database/dictionary of 3D skeletons for nearest-neighbor look-up. Tekin *et al*. [42] fused 2D and 3D image cues re-

lying on 2D-3D correspondences. Wang *et al*. [46] use the 3D ground truth to train an intermediate ranking network to extract the depth ordering of pairwise human joints from a single RGB image. Sun *et al*. [41] use a 3D regression based on bone segments derived from joint locations as opposed to directly using joint locations. Since these methods model 2D to 3D mappings from a given dataset, they implicitly incorporate dataset-specific parameters such as camera projection matrices, distance of skeleton from the camera, and scale of skeletons. This enables these models to predict metric position of joints in 3D on similar datasets, but requires paired 2D-3D correspondences which are difficult to obtain.

Weakly Supervised: Approaches such as [3, 10, 44, 53, 54, 55] do not explicitly use paired 2D-3D correspondences, but use *unpaired* 3D data to learn priors on shape (3D basis) or pose (articulation priors). For example, Zhou *et al*. [54] use a 3D pose dictionary to learn pose priors and Brau *et al*. [3] employ an independently trained network that learns a prior distribution over 3D poses (kinematic and self-intersection priors). Tome *et al*. [44], Wu *et al*. [48] and Tung *et al*. [11] pre-train low-dimensional representations from 3D annotations to obtain priors for plausible 3D poses. Another form of weak supervision is employed by Ronchi *et al*. [39], where they train a network using relative depth ordering of joints to predict 3D pose from images. Dabral *et al*. [8] uses supervision of 3D skeletons in conjunction with anatomical losses based on joint angle limits and limb symmetry. Rhodin *et al*. [37] train via 2D data, using multiple images of a single pose in addition to supervision in using 3D data when available. An adversarial training paradigm was used by Yang *et al*. [50] to improve an existing 3D pose estimation framework, lifting in-the-wild images with no 3D ground truth and comparing them to existing 3D skeletons.

Similar to our work, the weakly supervised approach of Drover *et al*. [9] also makes use of 2D projections to learn a 3D prior on human pose. However, Drover *et al*. utilize the ground-truth 3D points to generate a large amount (12M) of synthetic 2D joints for training, thus augmenting the original 1.5M 2D poses in Human3.6M by almost 10 times. This allows them to synthetically over-sample the space of camera variations/angles to learn the 3D priors from those poses. In contrast, we do not use any ground truth 3D projection or 3D data in any form. The fact that we can utilize multiple 2D datasets without any 3D supervision sets us apart from these previous approaches, and enables our method to exploit the large amount of available 2D pose data.

Unsupervised: Recently, Rhodin *et al*. [36] proposed an unsupervised method to learn a geometry-aware body representation. Their approach maps one view of the human to another view from a set of given multi-view images. It relies on synchronized multi-view images of subjects to learn an encoding of scene geometry and pose. It also uses video sequences to observe the same subject at multiple time instants to learn appearance. In contrast, we do not require multi-view images or the ability to capture the same pose at multiple time instants. We learn 3D pose from 2D projections alone. Kudo *et al*. [23] present 3D error results (130.9 mm) that are comparable to the trivial baseline reported in [9] (127.3 mm).

Learning Using Adversarial Loss: Generative adversarial learning has emerged as a powerful framework for modeling complex data distributions, some use it to learn generative models [13, 15, 56], and [45] leverages it to synthesize hard examples, *etc*. Previous approaches have used adversarial loss for human pose estimation by using a discriminator to differentiate real/fake 2D poses [6] and real/fake 3D poses [11, 21]. To estimate 3D, these techniques still require 3D data or use a prior 3D pose models. In contrast, our approach applies an adversarial loss over randomly projected 2D poses of the generated 3D skeletons. Previous works on image-to-image translation such as CycleGAN [56] or CyCADA [15] also rely on a cycle consistency loss in the image domain to enable unsupervised training. However, we use geometric self-consistency and utilize consistency loss in 3D and 2D joint locations, resulting in a novel method for lifting.

3. Unsupervised 2D-3D Lifting

In this section, we describe our unsupervised learning approach to lift 2D pose to a 3D skeleton. Let $\mathbf{x}_i = (x_i, y_i), i = 1 \dots N$, denote N 2D pose landmarks of a skeleton with the root joint (midpoint between hip joints) located at the origin. Let $\mathbf{X}_i$ denote the corresponding 3D joint for each 2D joint. We assume a camera with unit focal length centered at the origin $(0, 0, 0)$. Note that because of the fundamental perspective ambiguity, absolute metric depths cannot be obtained from a single view. Therefore, we fix the distance of the skeleton to the camera to a constant c units. In addition, we normalize the 2D skeletons such that the mean distance from the head joint to the root joint is $\frac{1}{c}$ units in 2D. This ensures that 3D skeleton will be generated with a scale of $\approx$ 1 unit (head to root joint distance).

3.1. Lifting Network

The lifting network $\mathrm{G}(x)$ is a neural network that outputs the 3D joint for each 2D joint.

$$\mathrm{G}_{\theta_G}(\mathbf{x}) = \mathbf{X}, \tag{1}$$

where θ_G are the parameters of the lifter learned during training. Internally, the lifter estimates the depth offset d_i of each joint relative to the fixed plane at c units. The 3D

joint is computed as $\mathbf{X}_i = (x_i z_i, y_i z_i, z_i)$, where

$$z_i = \max(1, c + d_i). \tag{2}$$

3.2. Random Projections

The generated 3D skeletons are projected to 2D using random camera orientations and these 2D poses are sent to the lifter and discriminator. Let $\mathbf{R}$ be a random rotation matrix, created by uniformly sampling an azimuth angle between [-π, π] and an elevation angle between [-π/9, π/9], and $\mathbf{X}_r$ be the location of the root joint of the generated skeleton. The rotated 3D skeleton $\mathbf{Y}_i$ is obtained as

$$\mathbf{Y}_i = Q(\mathbf{X}_i) = \mathbf{R} * (\mathbf{X}_i - \mathbf{X}_r) + \mathbf{T}, \tag{3}$$

where $\mathbf{T} = [0, 0, c]$. Q represents the rigid transformation between $\mathbf{Y}$ and $\mathbf{X}$. The rotated 3D skeleton $\mathbf{Y}_i$ is then projected to create a 2D skeleton $\mathbf{y}_i = P(\mathbf{Y}_i)$, where P denotes perspective projection.

3.3. Self-Supervision via Loop Closure

We now describe the symmetrical lifting and projection step performed on the synthesized 2D pose, $\mathbf{y}_i$. As shown in Figure 2, we lift the randomly projected pose $\mathbf{y}_i$ to obtain $\tilde{\mathbf{Y}}_i$

$$\tilde{\mathbf{Y}}_i = \mathrm{G}_{\theta_G}(\mathbf{y}_i). \tag{4}$$

$\tilde{\mathbf{Y}}_i$ is transformed to $\tilde{\mathbf{X}}_i$ by applying the inverse of rigid transformation Q that was used while generating the random projection $\mathbf{y}_i$ from $\mathbf{X}_i$. The 3D skeleton $\tilde{\mathbf{X}}_i$ is finally projected to the 2D skeleton $\tilde{\mathbf{x}}_i$.

Note that the lifting network $G(\cdot)$ remains the same in both the forward and backward part of the cycle as illustrated in Figure 2. If the lifting network accurately reconstructs the 3D pose from 2D inputs, then the 3D skeletons $\mathbf{Y}_i$ and $\tilde{\mathbf{Y}}_i$ and the corresponding 2D projections $\mathbf{x}_i$ and $\tilde{\mathbf{x}}_i$ should be similar. The cycle described herein provides a strong signal for self-supervision for the lifting network, whose loss term can be updated by adding two additional components, namely, $\mathcal{L}_{3D} = \left\|\mathbf{Y} - \tilde{\mathbf{Y}}\right\|^2$ and $\mathcal{L}_{2D} = \|\mathbf{x} - \tilde{\mathbf{x}}\|^2$.

3.4. Discriminator for 2D Poses

The 2D pose discriminator D is a neural network (with parameters θ_D) that takes as input a 2D pose and outputs a probability between 0 and 1. It classifies between real 2D pose $\mathbf{r}$ (target probability of 1) and fake (projected) 2D pose $\mathbf{y}$ (target probability of 0). Note that for any training sample $\mathbf{x}$ for lifter, we do not require $\mathbf{r}$ to be same as $\mathbf{x}$ or any of it's multi-view correspondences. During learning we utilize a standard GAN loss [13] defined as

$$\min_{\theta_G} \max_{\theta_D} \mathcal{L}_{adv} = \mathbb{E}(\log(D(\mathbf{r}))) + \mathbb{E}(\log(1 - D(\mathbf{y}))). \tag{5}$$

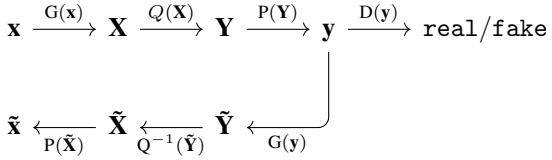

Figure 2. Self-supervision achieved by closing the loop between the generated skeleton $\mathbf{Y}$, its random projection $\mathbf{y}$. The recovered 3D skeleton $\tilde{\mathbf{Y}}$ is obtained by lifting $\mathbf{y}$. Upon reversing the geometric transformations, training can be self-supervised by comparing $\mathbf{x}$ with $\tilde{\mathbf{x}}$, and $\mathbf{Y}$ with $\tilde{\mathbf{Y}}$.

The discriminator provides feedback to the lifter allowing it to learn priors on 3D skeletons such as the ratio of limb lengths and joint angles using only random 2D projections, thus allowing it to avoid inadequacies as shown in Sect. 4.3.

3.5. Temporal Consistency

Note that our approach does not require video data for training. However, when available, temporal 2D pose sequences (*e.g.* video sequence of actions) can improve the accuracy of the single frame lifting network. We exploit the temporal smoothness via an additional loss function to refine the lifting network $G(\cdot)$ as shown in Figure 3. We train an additional discriminator, $T(\cdot)$ that takes as input the difference of 2D poses adjacent in time. The real data for this discriminator comes from a sequence of real 2D poses available during training, $\mathbf{r}_t - \mathbf{r}_{t+1}$. The discriminator $T(\cdot)$ is updated to optimize the loss that can distinguish the distribution of real 2D pose differences from those of the fake 2D (sequential) projections $\mathbf{y}_t - \mathbf{y}_{t+1}$. Specifically,

$$\begin{aligned}\max_{\theta_T} \mathcal{L}_T = &\mathbb{E}(\log(T(\mathbf{r}_t - \mathbf{r}_{t+1}))) + \\ &\mathbb{E}(\log(1 - T(\mathbf{y}_t - \mathbf{y}_{t+1}))).\end{aligned} \tag{6}$$

3.6. Learning from 2D Poses in the Wild

To improve the 3D lifting accuracy in the target domain of interest (*e.g.* Human 3.6M, $\mathbf{x}_t$), we wish to augment 2D training data from in the wild (*e.g.* OpenPose joint estimates on Kinetics dataset, $\mathbf{x}_s$). Depending on the choice of 2D pose extraction algorithms [4, 30, 47], the position and semantics of 2D keypoints can vary greatly from the representation adopted by the target domain (*e.g.* center of face vs. top of the head or side of the hips vs. pelvis).

We train a 2D domain adapter neural network C to map the source domain 2D joints to target domain 2D joints (see Figure 4). Let $\mathbf{x}_{sc}$ denote the corrected source domain 2D joints, such that $\mathbf{x}_{sc} = \mathbf{x}_s + C(\mathbf{x}_s)$. Note that we do not assume any correspondences between the 2D joints in the source and target domains. Thus, we cannot train C using any form of supervised loss. In absence of any supervision,

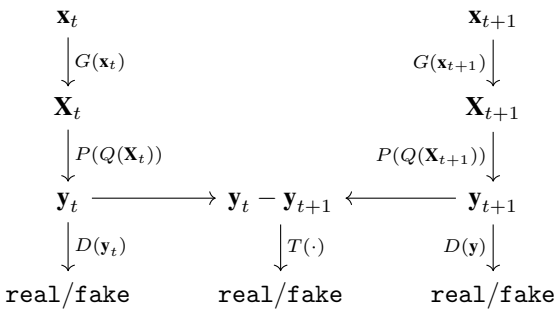

Figure 3. Discriminator $T(\cdot)$ enforces a distribution on the temporal differences of projected 2D poses. The temporal consistency is an optional element to stabilize the results, and is only added during training, allowing inference on single frame 2D pose inputs. Subscripts t and $t+1$ denote two consecutive inputs. Lifting, transformation and projection is done as in Figure 2.

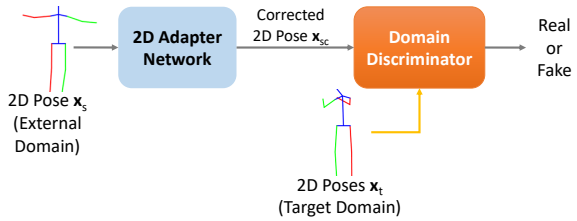

Figure 4. Unsupervised domain adaptation to transform 2D poses from source domain to match the semantics of the target domain. The 2D adapter modifies the input $\mathbf{x}_s$ to generate corrected pose $\mathbf{x}_{sc}$. The domain discriminator is used to ensure that the distribution of adapted2D poses $\mathbf{x}_{sc}$ and the target domain 2D poses $\mathbf{x}_t$ match. Adapted poses $\mathbf{x}_{sc}$ are used for training a domain agnostic lifter as shown in Figure 1.

we use a domain discriminator D_D to match the distribution between the two domains. Again utilizing the standard GAN loss [13], we optimize the following loss

$$\min_{\theta_C} \max_{\theta_{D_D}} \mathcal{L}_{adv} = \mathbb{E}(\log(D_D(\mathbf{x}_p))) + \lambda \left\| C(\mathbf{x}_s) \right\|^2 + \mathbb{E}(\log(1 - D_D(\mathbf{x}_{sc}))), \quad (7)$$

where, the $\lambda \left\| C(\mathbf{x}_s) \right\|^2$ is a regularizer term to keep the corrections limited to a small magnitude.

Figure 5 shows an example of the difference in semantics between the Human3.6M (target domain) and OpenPose (source domain). In OpenPose, the top of the head is not marked and the center of the marked eye joints are used. In addition, the shoulder keypoints are marked higher than in Human3.6M. The domain adapted 2D pose (middle) is closer in terms of keypoint locations to the target domain. Our domain correction is an off-line preprocessing step. The domain corrected 2D poses, $\mathbf{x}_{sc}$, are fed both to the lifting network (Sect. 3.1) and the 2D pose discriminator (Sect. 3.4) during training.

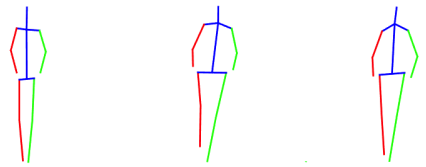

Figure 5. An example of unsupervised domain adaptation. (Left) 2D joints estimated using OpenPose for an example DeepMind Kinetics image. (Middle) Resulting 2D pose after adaptation. (Right) Similar pose from Human3.6M dataset. Notice the change in the width of hips and slant of shoulders after adaptation.

3.7. Training

As discussed, the 2D to 3D lifting network is trained using geometric self-supervision along with 2D pose and temporal discriminators. Network parameters are updated to optimize the total loss given by,

$$\mathcal{L} = \mathcal{L}_{adv} + w_{2D}\mathcal{L}_{2D} + w_{3D}\mathcal{L}_{3D} + w_T\mathcal{L}_T, \quad (8)$$

where, $w_{2D} = 10, w_{3D} = 0.001$, and $w_T = 1$ are the relative weights for each of the 2D, 3D, and temporal loss terms, respectively.

Architectures: We do not use any convolutional layers and use fully connected layers (followed by residual blocks) for all the neural networks described above. Both the lifting network and the 2D pose discriminator takes as input $2N$ dimensional vectors, where N denotes the number of 2D/3D pose points. Similarly, the temporal discriminator takes $2N + 2NM$ inputs corresponding to the pose joints in the current frame and their temporal differences with M other consecutive frames (before and/or after). We adopt a similar architecture as that of Martinez *et al*. [26], with the lifting network composed of 4 residual blocks and the discriminator with 3 residual blocks. For 2D domain adaptation, we use 4 residual blocks for the adapter and 3 residual blocks for the domain discriminator. Batch normalization was used in the lifter and the adapter but not in either of the discriminators. Our experiments use $N = 14$ joint locations. For training we used a batch size of 8192, a constant depth $c = 10$ and the Adam optimizer.

4. Experimental Evaluation

We present quantitative and qualitative results on the widely used Human3.6M dataset [18] for benchmarking. Additionally, to demonstrate how the unsupervised learning framework can be improved by leveraging 2D pose data from in-the-wild images, we augment our training data by adapting OpenPose estimated 2D human poses from the Kinetics [22] dataset. We also show qualitative visualizations of reconstructed 3D skeletons from 2D pose landmarks on the MPII [2] and Leeds Sports Pose (LSP) [20] datasets, for which the ground truth 3D data is not available.

4.1. Dataset and Metrics

Human3.6M Dataset: Human3.6M is one of the largest 3D human pose datasets, consisting of 3.6 million 3D human poses. The dataset contains video and motion capture (MoCap) data from 5 female and 6 male subjects. Data is captured from 4 different viewpoints, while subjects perform typical activities such as talking on phone, walking, eating, *etc*.

MPI-INF-3DHP: The MPI-INF-3DHP [27] is a large human pose dataset containing >1.3M frames taken from diverse viewpoints. The dataset has 4 male and 4 female actors performing an array of actions similar to but more diverse than the Human3.6M dataset.

Kinetics dataset: The Kinetics dataset contains 400 video clips each for 400 activities involving one or more persons. The video clips are sourced from Youtube and each clip is approximately 10 seconds in duration. We did not use any of the class annotations from the dataset for our training. Instead, we extracted 2D pose landmarks using OpenPose [4] on sampled frames from this dataset. We retained only those frames in which all the landmarks on a person were estimated with sufficient confidence. After this filtering, approximately 9 million 2D skeletons were obtained.

Evaluation Metric: We report the Mean Per Joint Position Error (MPJPE) in millimeters after scaling and rigid alignment to the ground truth skeleton. Similar to previous works [9, 11, 24, 26, 36, 43, 54], we report results on subjects S9 and S11. Also, following the convention as in [26, 36], we only use data from subjects S1, S5, S6, S7, and S8 for training. We do not train class specific models or leverage any motion information during inference to improve the results. The reported metrics are taken from the respective papers for comparisons. We also compare our method to [27, 53] which uses the adapted Percentage of Correct Keypoints (PCK) and corresponding Area Under Curve (AUC) metrics.

4.2. Quantitative Results

We summarize our results for Human3.6M and MPI-INF-3DHP in Table 1 and Table 2, respectively. In addition to comparing with the state-of-the-art unsupervised 3D pose estimation method of Rhodin *et al*. [36], we also show results from top fully supervised and weakly supervised methods. Results from [36] uses images as input and are hence comparable to Ours(SH) results which use 2D joints extracted from the same input images using SH detector [30]. Our method reduces error by 30% compared to [36] (68mm vs. 98.2mm).

Table 3 shows the results of an ablation study on lifter with various algorithmic components using ground truth 2D points. **SS** denotes self-consistency (Sect. 3.3), **Adv** adds the 2D pose discriminator (Sect. 3.4), **DA** augments the training data by adapting 2D poses from Kinetics (Sect. 3.6), and **TD** leverages temporal cues during training (Sect. 3.5), when available. As further analyzed in Sect. 4.3, just using self consistency loss can lead to unrealistic skeletons without the additional discriminator. Augmenting our approach with additional 2D poses obtained from the Kinetics dataset (Ours: Adv + SS + DA) further reduces the error down to 55mm. Lastly, we exploit temporal informa-

Supervision	Algorithm	Error (mm) GT	Error (mm) IMG
Full	Chen *et al*. [5]	57.5	82.7
	Martinez *et al*. [26]	37.1	52.1
Weak	3DInterpreter [48]	88.6	98.4
	AIGN [11]	79.0	97.2
	Drover *et al*. [9]	38.2	64.6
Unsupervised	Rhodin *et al*. [36]	-	98.2
	Ours	**51**	**68**

Table 1. Comparison to the state-of-the-art unsupervised method of Rhodin *et al*. [36] on Human3.6M. Comparable metrics for fully/weakly supervised methods are included for reference. Our approach outperforms [36] and several weakly supervised approaches [11, 48] by a significant margin. GT and IMG denote results using ground truth 2D pose and estimated 2D pose by SH/CPM [30, 47], respectively.

Supervision	Algorithm	Trainset	PCK	AUC
Full	Mehta [27]	MPI	72.5	36.9
	Mehta [27]	H36M	64.7	31.7
Weak	Zhou [53]	H36M	69.2	32.5
Unsupervised	Ours	MPI	**71.1**	**36.3**
	Ours	H36M	**64.3**	**31.6**

Table 2. Our results (14-joint) on MPI-INF-3DHP with metrics as in Mehta *et al*. [27]. The proposed unsupervised approach achieves similar performance as [27] and [53].

Type	Ablation	Error (mm)
Architecture/ Loss Variations	SS	162
	SS + Symm	168
	Adv	61
	Adv+DA	59
	Adv+SS	58
	Adv+SS+DA	55
	Adv+SS+DA+TD	**51**
Supervised Fine-Tuning with 3D Data	0%	55
	5%	37

Table 3. Ablation studies.The architecture/loss ablations show the effect of various components on the unsupervised training. Supervised fine-tuning with only 5% of randomly sampled Human3.6M 3D data gives similar performance as compared to fully-supervised results of [26].

tion during training (Ours: Adv + SS + DA + TD), when available, to obtain an error of 51mm on Human3.6M. It should be noted that the inference for the TD experiment is still done on single frames and the results can be further improved by applying temporal smoothness techniques on video sequences.

4.3. Inadequacy of Geometric Self-Supervision

At first glance, it may appear that self supervision is sufficient to learn a good lifter, without the need for a discriminator. However, we found that in absence of the 2D pose discriminator, network can produce outputs which are geometrically self-consistent, but not realistic (see Figure 7). We present an analysis of the 3D outputs that the lifting network can generate with only self-supervision. Specifically, we examine the ratios of upper to lower arm and leg, both for the left and right side of human body (4 ratios).

Figure 6 (Left) shows the distribution of the 4 ratios, for a lifter trained using self-consistency loss alone. Note that the lifter produces different limb length ratios for the left and right side of the body. Thus self-consistency loss alone may not produce symmetric (realistic) skeletons without any 3D priors. Figure 6 (Middle) shows that after imposing symmetry constraints, the distributions of the left and right limbs are better aligned. However, the distributions are *flatter* since enforcing the *same* ratios for left and right sides does not ensure that these ratios are *realistic* (conforming to a human body). In other words, the lifter may choose different ratios for different training examples. Figure 6 (Right) shows the distributions when a discriminator that gives feedback to the lifter using real 2D poses is used. Notice that the ratio distributions become sharper and closer to distributions of real ratios in the training set. This is the reason that using self-supervision loss (SS) alone performs worse in our ablation studies as shown in Table 1. However, the self-consistency further improves the performance in conjunction with 2D pose discriminator (Adv+SS).

Note that we do not use symmetry ratios in our framework when the discriminator is present. Our lifting network can learn higher order 3D skeleton statistics (beyond symmetry) based on the feedback from geometric self consistency and the 2D pose discriminator.

4.4. Semi-supervised 3D Pose Estimation

Other methods have shown improvement in accuracy when a small amount of 3D data is used for supervised fine-tuning. We fine tuned our baseline model (from unsupervised training) using 5% of randomly sampled 3D data available in Human3.6M dataset. With this, our method could achieve performance comparable to fully supervised method (37mm) as shown in Table 3.

4.5. Qualitative Results

Figure 8 shows some of the 3D pose reconstruction results on Human3.6M dataset using our lifting network. The ground truth 3D skeleton is depicted in gray. Some of the failures are shown in Figure 9. Most of these can be attributed to self-occlusions or flip ambiguities in viewing direction (for more details see Suppl. materials).

To demonstrate generalization, we show some examples of 3D skeletons estimated on MPII [2] and the Leeds Sports Pose (LSP) [20] datasets, in Figures 10 and 11 respectively. MPII has images extracted from short Youtube videos. LSP dataset consists of images of sport activities sampled from Flickr. Our unsupervised method successfully recovers 3D poses on these datasets without being trained on them.

4.6. Discussion

Previous unsupervised and weakly supervised methods use additional constraints on training data in lieu of 3D annotations. For example, [9, 51] leverage synthetic 2D poses obtained from known 3D skeletons to improve results. Similarly, Rhodin *et al*. [36] derive an appearance and geometric model by choosing different frames from temporal sequences and multi-view images involving the same person. However, in theory, if multi-view images from synchronized cameras are available, one could triangulate the detected 2D joints to get 3D joints and train a supervised network. In contrast, our method treats each 2D skeleton as an individual training example, without requiring any multi-view correspondence. Hence, there is no restriction on where the 2D input pose originates; it could be obtained from a single image, video, or multi-view sequences. Our work explores the innate geometry of human pose itself, whereas [36] exploits the consistency in camera geometry and appearance of specific individuals. As shown in Sect. 4.2, our approach is able to augment the training data from other datasets (*e.g*. Kinetics) with 2D skeletons captured in the wild.

Our current approach cannot handle occluded/missing joints during training or testing phases. This limits the amount of external domain data that can be used for training. For example, using OpenPose on Kinetics dataset results in 17M skeletons with at least 10 joints, but only 9M complete skeletons (14 joints). Though not the main focus of the paper, we did a small experiment to fill-in missing joints to further augment our training data. We trained a two-layer fully connected neural network which takes incomplete OpenPose 2D pose estimates on Human3.6M images as input and outputs completed 14 joints. The network was trained using the corresponding 2D ground-truth joints from Human3.6M in a supervised manner. Using the completed poses (17M skeletons) from the Kinetic dataset, our method achieved a MPJPE of 48mm on Human3.6M test data. This experiment further underscores the importance

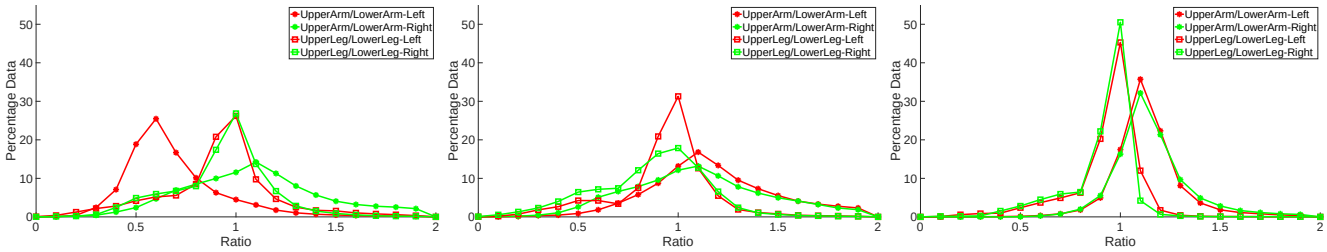

Figure 6. Distribution of limb lengths ratios on Human3.6M test data. (Left) Training using self-consistency loss alone does not impose symmetry. (Middle) Using self-consistency and symmetry aligns distribution of left/right limbs, but results in flatter (unrealistic) distributions. (Right) Using a discriminator sharpens the distributions and brings them closer to real values (ground truth ratio is ~ 1.0 and ~ 1.1 for leg and arms respectively.)

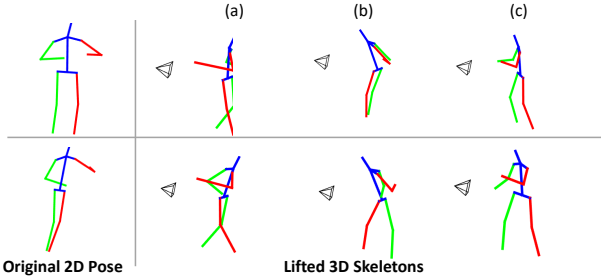

Figure 7. Inadequacy of self-consistency loss. Left most Col is input 2D pose. (a) Self-consistency alone is unable to recover the correct 3D skeletons. (b) With symmetry constraints, limb lengths become symmetric but may not have realistic ratios. (c) Adding 2D pose discriminator results in a geometrically consistent and realistic 3D skeletons.

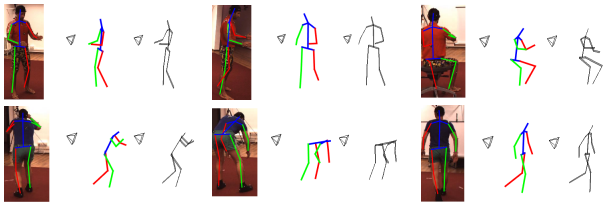

Figure 8. Qualitative results on Human3.6M dataset. (Left to right) Color image with overlaid 2D pose points, estimated and ground-truth 3D skeleton.

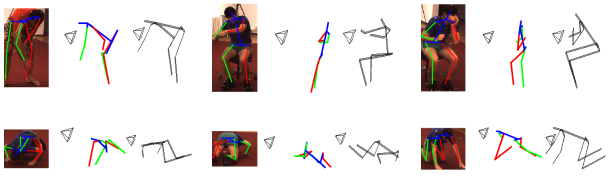

Figure 9. Examples of failure cases of our algorithm on Human3.6M dataset. Ground truth 3D skeletons are shown in gray.

of volume and diversity of training data for unsupervised learning. We believe that by using auto-encoders and other unsupervised methods for data completion will enable utilizing even more diverse datasets, where 2D joints may be extracted from a variety of 2D pose estimation algorithms. Future work includes training the filling network and the domain adaptation network together with the lifting network

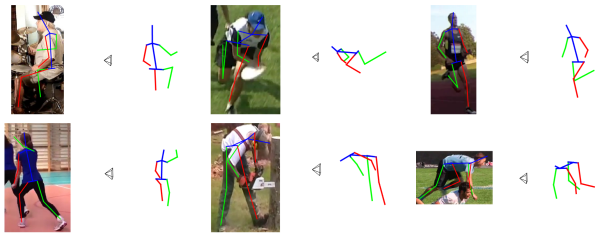

Figure 10. Examples of 3D pose reconstruction on images from MPII dataset (no ground-truth 3D skeleton). Each image shows overlaid 2D pose and the estimated 3D skeleton.

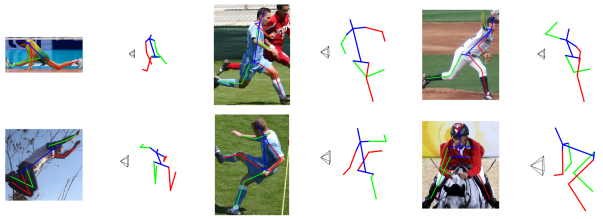

Figure 11. Examples of 3D pose reconstruction on images from LSP dataset (no ground-truth 3D).

in an end-to-end manner.

5. Conclusions

For 3D human pose estimation, acquiring 3D MoCap data remains an expensive and challenging endeavor. We presented an unsupervised learning approach to generate 3D skeletons from 2D joints, which does not require 3D data in any form. Our paper introduces geometric self-supervision as a novel constraint to learn the 2D-3D lifter. We showed that while geometric self-supervision is not a sufficient condition and cannot generate realistic skeletons by itself, it improves the reconstruction accuracy when combined with a discriminator. By training a domain adapter, we showed how to utilize data from different domains and datasets in an unsupervised manner. Thus, we believe that our paper has significantly improved the state-of-art in unsupervised learning of 3D skeletons by developing the key idea of geometric self-supervision and utilizing domain adaptation. Future work includes end-to-end training of 3D skeletons from 2D images, using self-supervision.

References

[1] Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele. Multi-view pictorial structures for 3d human pose estimation. In *BMVC*, 2013. 1

[2] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014. 5, 7

[3] Ernesto Brau and Hao Jiang. 3d human pose estimation via deep learning from 2d annotations. In *Fourth International Conference on 3D Vision*, 2016. 3

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2, 4, 6

[5] C.-H. Chen and D. Ramanan. 3d human pose estimation = 2d pose estimation + matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2, 6

[6] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial posenet: A structure-aware convolutional network for human pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 3

[7] Xiao Chu and Alan Yuille. Orinet: A fully convolutional network for 3d human pose estimation. In *BMVC*, 2018. 2

[8] Rishabh Dabral, Anurag Mundhada, Uday Kusupati, Safeer Afaque, Abhishek Sharma, and Arjun Jain. Learning 3d human pose from structure and motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 668–683, 2018. 3

[9] Dylan Drover, Rohith MV, Ching-Hang Chen, Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh. Can 3D pose be learned from 2D projections alone? In *European Conference of Computer Vision Workshop*, 2018. 2, 3, 6, 7

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

[11] 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 *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2, 3, 6

[12] David A Forsyth, Okan Arikan, and Leslie Ikemoto. *Computational Studies of Human Motion: Tracking and Motion Synthesis*. Now Publishers Inc, 2006. 1

[13] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, and S Ozair. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014. 3, 4, 5

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 2

[15] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei A. Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. *International Conference on Machine Learning (ICML)*, 2018. 3

[16] Michael Hofmann and Dariu M Gavrila. Multi-view 3d human pose estimation in complex environment. *IJCV*, 2012. 1

[17] David Hogg. Model-based vision: a program to see a walking person. *Image and Vision computing*, 1983. 1

[18] 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 Trans. Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, Jul 2014. 5

[19] Hao Jiang. 3d human pose reconstruction using millions of exemplars. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010. 2

[20] Sam Johnson and Mark Everingham. Clustered pose and nonlinear appearance models for human pose estimation. In *Proceedings of the British Machine Vision Conference*, pages 12.1–12.11. BMVA Press, 2010. doi:10.5244/C.24.12. 5, 7

[21] Angjoo Kanazawa, Michael Black, David Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[22] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Apostol Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 5

[23] Yasunori Kudo, Keisuke Ogaki, Yusuke Matsui, and Yuri Odagiri. Unsupervised adversarial learning of 3d human pose from 2d joint locations. *arXiv preprint arXiv:1803.08244*, 2018. 3

[24] Sijin Li and Antoni B Chan. 3d human pose estimation from monocular images with deep convolutional neural network. In *ACCV*, 2014. 6

[25] Sijin Li, Weichen Zhang, and Antoni B. Chan. Maximum-margin structured learning with deep networks for 3d human pose estimation. In *IEEE Inter-*

national Conference on Computer Vision (ICCV), December 2015. 2

[26] Julieta Martinez, Rayat Hossain, Javier Romero, and James Little. A simple yet effective baseline for 3d human pose estimation. In *ICCV*, 2017. 2, 5, 6

[27] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *3D Vision (3DV), 2017 Fifth International Conference on*. IEEE, 2017. 2, 6

[28] Thomas B Moeslund and Erik Granum. A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding*, 2001. 1

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

[30] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pages 483–499. Springer, 2016. 2, 4, 6

[31] Joseph O'Rourke and Norman I Badler. Model-based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1980. 1

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

[33] Sungheon Park and Nojun Kwak. 3d human pose estimation with relational networks. In *BMVC*, 2018. 2

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

[35] Umer Rafi, Juergen Gall, and Bastian Leibe. A semantic occlusion model for human pose estimation from a single depth image. In *Proceedings of the IEEE Conference on CVPR Workshops*, 2015. 1

[36] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3D human pose estimation. In *European Conference of Computer Vision*, 2018. 3, 6, 7

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

[38] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 2

[39] 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. 3

[40] Jamie Shotton, Toby Sharp, Alex Kipman, Andrew Fitzgibbon, Mark Finocchio, Andrew Blake, Mat Cook, and Richard Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013. 1

[41] 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. 3

[42] Bugra Tekin, Pablo Marquez-Neila, Mathieu Salzmann, and Pascal Fua. Learning to fuse 2d and 3d image cues for monocular body pose estimation. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[43] Bugra Tekin, Artem Rozantsev, Vincent Lepetit, and Pascal Fua. Direct prediction of 3d body poses from motion compensated sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 991–1000, 2016. 1, 6

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

[45] Shashank Tripathi, Siddhartha Chandra, Amit Agrawal, Ambrish Tyagi, James M. Rehg, and Visesh Chari. Learning to generate synthetic data via compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 3

[46] Min Wang, Xipeng Chen, Wentao Liu, Chen Qian, Liang Lin, and Lizhuang Ma. Drpose3d: Depth ranking in 3d human pose estimation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 978–984, 2018. 3

[47] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, pages 4724–4732, 2016. 2, 4, 6

[48] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. Single image 3d interpreter network. In *ECCV*, pages 365–382, 2016. 3, 6

[49] Bruce Xiaohan Nie, Ping Wei, and Song-Chun Zhu. Monocular 3d human pose estimation by predicting depth on joints. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[50] Wei Yang, Wanli Ouyang, Xiaolong Wang, Jimmy S. J. Ren, Hongsheng Li, and Xiaogang Wang. 3D human pose estimation in the wild by adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2018. 3

[51] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2, 7

[52] Ho Yub Jung, Soochahn Lee, Yong Seok Heo, and Il Dong Yun. Random tree walk toward instantaneous 3d human pose estimation. In *CVPR*, 2015. 1

[53] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *IEEE International Conference on Computer Vision*, 2017. 3, 6

[54] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. Sparseness meets deepness: 3d human pose estimation from monocular video. In *CVPR*, 2016. 1, 3, 6

[55] Xiaowei Zhou, Menglong Zhu, Georgios Pavlakos, Spyridon Leonardos, Kostantinos G Derpanis, and Kostas Daniilidis. Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 3

[56] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3