Can 3D Pose be Learned from 2D Projections Alone?

Dylan Drover, Rohith MV, Ching-Hang Chen,
Amit Agrawal, Ambrish Tyagi, and Cong Phuoc Huynh

Amazon Lab126 Inc., Sunnyvale, CA, USA
{droverd, kurohith, chinghc, aaagrawa,
ambrisht, conghuyn}@amazon.com

Abstract. 3D pose estimation from a single image is a challenging task in computer vision. We present a weakly supervised approach to estimate 3D pose points, given only 2D pose landmarks. Our method does not require correspondences between 2D and 3D points to build explicit 3D priors. We utilize an adversarial framework to impose a prior on the 3D structure, learned solely from their random 2D projections. Given a set of 2D pose landmarks, the generator network hypothesizes their depths to obtain a 3D skeleton. We propose a novel Random Projection layer, which randomly projects the generated 3D skeleton and sends the resulting 2D pose to the discriminator. The discriminator improves by discriminating between the generated poses and pose samples from a real distribution of 2D poses. Training does not require correspondence between the 2D inputs to either the generator or the discriminator. We apply our approach to the task of 3D human pose estimation. Results on Human3.6M dataset demonstrates that our approach outperforms many previous supervised and weakly supervised approaches.

Keywords: Weakly Supervised Learning, Generative Adversarial Networks, 3D Pose Estimation, Projective Geometry

1 Introduction

Inferring 3D human poses from images and videos (automatic motion-capture) has garnered particular attention in the field [15, 32, 11, 29] due to its numerous applications related to tracking, action understanding, human-robot-interaction and gaming, among others. Estimating 3D pose of articulated objects from 2D views is one of the long-standing ill-posed inverse problems in computer vision. We have access to, and continue to generate, large amounts of image and video data at an unprecedented rate. This begs the question: Can we build a system that can estimate the 3D joint locations/skeleton of humans by leveraging this abundant 2D image and video data?

The problem of training end-to-end, image to 3D, pose estimation models is challenging due to variations in background, illumination, appearance, camera

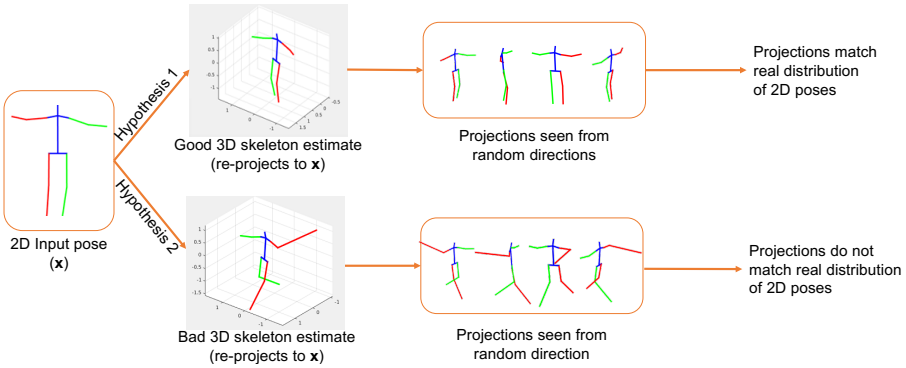

Fig. 1. Key intuition behind our approach: A generator can hypothesize multiple 3D skeletons for a given input 2D pose. However, only plausible 3D skeletons will project to realistic looking 2D poses after random projections. The discriminator evaluates the "realness" of the projected 2D poses and provides appropriate feedback to the generator to learn to produce realistic 3D skeletons

characteristics, *etc*. Recent approaches [26, 30] have decomposed the 3D pose estimation problem into (i) estimating 2D landmark locations (corresponding to skeleton joints) and (ii) estimating 3D pose from them (lifting 2D points to 3D). Following such a scheme, suitable 2D pose estimators can be chosen based on the application domain [42, 31, 6, 13] to estimate 2D poses, which can then be fed to a common 2D-3D lifting algorithm for recovering 3D pose.

A single 2D observation of landmarks admits infinite 3D skeletons as solution; not all these are physically plausible. The restriction of solution space to realistic poses is typically achieved by regularizing the 3D structure using priors such as symmetry, ratio of length of various skeleton elements, and kinematic constraints. These priors are often learned from ground truth 3D data, which is limited due to the complexity of capture systems. We believe that leveraging unsupervised algorithms such as generative adversarial networks for 3D pose estimation will help address the limitations of capturing such 3D data. Our work addresses the fundamental problem of lifting 2D image coordinates to 3D space without the use of any additional cues such as video [47, 40], multi-view cameras [2, 14], or depth images [35, 46, 38].

We present a weakly supervised learning algorithm to estimate 3D human skeleton from 2D pose landmarks. Unlike previous approaches we do not learn priors explicitly through 3D data or utilize explicit 2D-3D correspondence. Our system can generate 3D skeletons by only observing 2D poses. Our paper makes the following contributions:

- We present and demonstrate that a latent 3D pose distribution can be learned solely by observing 2D poses, without requiring any regression from 3D data.
- We propose a novel Random Projection layer and utilize it along with adversarial training to enforce a prior on 3D structure from 2D projections.

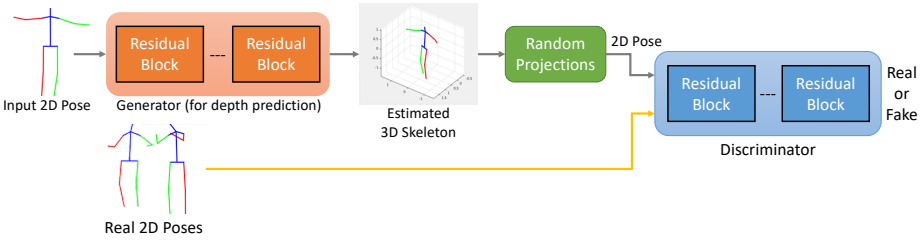

Fig. 2. Adversarial training architecture for learning 3D pose

Figure 1 outlines the key intuition behind our approach. Given an input 2D pose, there are an infinite number of 3D configurations whose projections match the position of 2D landmarks in that view. However, it is very unlikely that an implausible 3D skeleton looks realistic from another randomly selected viewpoint. On the contrary, random 2D projections of accurately estimated 3D poses are more likely to conform to the real 2D pose distribution, regardless of the viewing direction. We exploit this property to learn the prior on 3D via 2D projections. For a 3D pose estimate to be accurate (a) the projection of the 3D pose onto the original camera should be close to the detected 2D landmarks, and (b) the projection of the 3D pose onto a random camera should produce 2D landmarks that fit the distribution of real 2D landmarks.

Generative Adversarial Networks (GAN) [12] provide a natural framework to learn distributions without explicit supervision. Our approach learns a latent distribution (3D pose priors) indirectly via 2D poses. Given a 2D pose, the generator hypothesizes the relative depth of joint locations to obtain a 3D human skeleton. Random 2D projections of the generated 3D skeleton are fed to the discriminator, along with actual 2D pose samples (see Figure 2). The 2D poses fed to the generator and discriminator do not require any correspondence during training. The discriminator learns a prior from 2D projections and enables the generator to eventually produce realistic 3D skeletons.

We demonstrate the effectiveness of our approach by evaluating 3D pose estimation on the Human3.6M dataset [17]. Our method shows an improvement over other weakly supervised methods which use 2D pose as input [10, 43]. Interestingly we also outperform a number of supervised methods that use explicit 2D-3D correspondences.

The remainder of this paper is organized as follows. We discuss related work in Section 2. Section 3 provides details of the proposed method and the training methodology. Our experimental evaluation results are presented in Section 4. Finally, we close with concluding remarks in Section 5.

2 Related Work

2D Pose Estimation: Significant progress has been made recently in 2D pose estimation using deep learning techniques [42, 31, 6, 13]. Newell *et al.* [31] pro-

posed a stacked hourglass architecture for predicting heatmap of each 2D joint location. Convolutional Pose Machines (CPM) [42] employ a sequential architecture by combining the prediction of previous stages with the input image to produce increasingly refined estimates of part locations. Cao *et al.* [6] also estimate a part affinity field along with landmark probabilities and uses a fast, greedy search for real-time multi-person 2D pose estimation. Kaiming *et al.* [13] accomplish this by performing fine-grained detection on top of an object detector. Our proposed method will continue to benefit from the ongoing improvement of 2D pose estimation algorithms.

3D Pose Estimation: Several approaches try to directly estimate 3D joint locations from images [34, 33, 37, 49, 28] in an end-to-end learning framework. However, in this paper we focus on benchmarking against methods which estimate 3D pose from 2D landmark positions, known as lifting from 2D to 3D [26, 10, 7]. Since the input to the methods is only 2D landmark locations, it is easy to augment training of these methods using synthetic data. Like other methods in this category, our method could be combined with a variety of 2D pose estimators based on the application without retraining. To better distinguish our work from previous approaches on lifting 2D pose landmarks to 3D, we define the following categories:

Fully Supervised: These include approaches such as [26, 44, 23] that use paired 2D-3D data comprised of ground truth 2D locations of joint landmarks and 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 [30] learns a regression from a 2D distance matrix to a 3D distance matrix using 2D-3D correspondences. Exemplar based methods [7, 45, 18] use a database/dictionary of 3D skeletons for nearest-neighbor look-up. Lin *et al.* [24] learn an end-to-end Recurrent Pose Sequence Machine whereas our approach does not use any video information. Mehta *et al.* [28] combine a regression network which estimates 2D and 3D poses with, temporal smoothing and a parameterized, kinematic skeleton fitting method to produce stable 3D skeletons across time. Tekin *et al.* [39] fuse 2D and 3D image cues relying on 2D-3D correspondences. Since these methods model 3D mapping from a given dataset, they implicitly incorporate dataset-specific parameters such as camera projection matrices, distance of skeleton from camera, and scale of skeletons. This enables such 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 [47, 41, 9, 5] use *unpaired* 3D data to learn a prior, typically as a 3D basis or articulation priors, but do not explicitly use paired 2D-3D correspondences. For example, Tome *et al.* [41] pre-train a low-rank Gaussian model from 3D annotations as a prior for plausible 3D poses. Wu *et al.* [43] proposed a 3D interpreter network that also estimates the weights of a 3D basis, which are learned separately for each object class using 3D data. Similarly Tung *et al.* [10] build a 3D shape basis using PCA by aligning 3D skeletons and predicting basis coefficients. Though this method accepts 2D landmark locations as input, this information is represented as an image

within the network. On the other hand, we directly operate on the vectors of 2D landmark pixel locations, with the advantage of working in lower dimensions and avoiding convolution layers in our network. Zhou *et al.* [47] also use a 3D pose dictionary to learn pose priors. Brau *et al.* [5] employ an independently trained network that learns a prior distribution over 3D poses (kinematic and self-intersection priors) to impose constraints. Zhou *et al.* [49] combine the 2D pose estimation task with a constraint on the bone length ratio in each skeleton group for image-to-3D pose estimation.

Learning Using Adversarial Loss: Recently, Generative Adversarial Networks (GAN) [12] have emerged as a powerful framework for learning generative models for complex data distributions. In a GAN framework, a generator is trained to synthesize samples from a latent distribution and a discriminator network is used to distinguish between synthetic and real samples. The generator's goal is to fool the discriminator by producing samples that match the distribution of real data. Previous approaches have used adversarial loss for human pose estimation by using a discriminator to differentiate real/fake 2D poses [8] and real/fake 3D poses [10, 20]. To estimate 3D, these techniques still require 3D data or use prior 3D pose model. Kanazawa *et al.* [20] trained an end-to-end system to estimate a skinned multi-person linear model (SMPL) [25] 3D mesh from RGB images by minimizing the re-projection error between 3D and 2D landmarks. However, they impose a prior on 3D skeletons using an adversarial loss with a large database of 3D human meshes. In contrast, our approach applies an adversarial loss over randomly projected 2D poses of the hypothesized 3D poses.

3 Weakly supervised Lifting of 2D Pose to 3D Skeleton

In this section we describe our weakly supervised learning approach to lift 2D human pose points to a 3D skeleton. Adversarial networks are notoriously difficult to train and we discuss design choices that lead to stable training. For consistency with generative adversarial network naming conventions, we will refer to the 3D pose estimation network as a generator. For simplicity, we work in the camera coordinate system, where the camera with unit focal length is centered at the origin $(0, 0, 0)$ of the world coordinate system. Let $\mathbf{x}_i = (x_i, y_i), i = 1 \dots N$, denote N 2D pose landmarks with the root joint (midpoint between hip joints) located at the origin. The 2D input pose is hence denoted by $\mathbf{x} = [\mathbf{x}_1 \dots \mathbf{x}_N]$. For numerical stability, we aim to generate 3D skeletons such that the distance from the top of the head to the root joint is approximately 1 unit.

Generator: The generator G is defined as a neural network that outputs a *depth offset* o_i for each point $\mathbf{x}_i$

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

where θ_G are parameters of the generator learned during training. The depth of each point is defined as

$$z_i = \max(0, d + o_i) + 1, \tag{2}$$

where d denotes the distance between the camera and the 3D skeleton. Note that the choice of d is arbitrary provided that $d > 1$. Constraining z_i to be greater than 1 ensures that the points are projected in front of the camera. In practice we use $d = 10$ units.

Next, we define the back projection and the random projection layers responsible for generating the 3D skeleton and projecting it to other random views.

Back Projection Layer: The back projection layer takes the input 2D points $\mathbf{x}_i$ and the predicted z_i to compute a 3D point $\mathbf{X}_i = [z_i x_i, z_i y_i, z_i]$. Note that we use exact perspective projection instead of approximations such as orthographic or paraperspective projection.

Random Projection Layer: The hypothesized (generated) 3D skeleton is projected to 2D poses using randomly generated camera orientations, to be fed to the discriminator. For simplicity, we randomly rotate the 3D points (in-place) and apply perspective projection to obtain *fake* 2D projections. Let $\mathbf{R}$ be a random rotation matrix and $\mathbf{T} = [0, 0, d]$. Let $\mathbf{P}_i = [P_i^x, P_i^y, P_i^z] = \mathbf{R}(\mathbf{X}_i - \mathbf{T}) + \mathbf{T}$ denote the 3D points after applying the random rotation. These points are re-projected to obtain fake 2D points $\mathbf{p}_i = [p_i^x, p_i^y] = [P_i^x / P_i^z, P_i^y / P_i^z]$. The rotated points $\mathbf{P}_i$ should also be in front of the camera. To ensure that, we also force $P_i^z \geq 1$. Let $\mathbf{p} = [\mathbf{p}_1 \dots \mathbf{p}_N]$ denote the 2D projected pose.

Note that there is an inherent ambiguity in perspective projection; doubling the size of the 3D skeleton and the distance from the camera will result in the same 2D projection. Thus a generator that predicts absolute 3D coordinates has an additional degree of freedom between the predicted size and distance for each training sample in a batch. This could potentially result in large variance in the generator output and gradient magnitudes within a batch and cause convergence issues in training. We remove this ambiguity by predicting depth offsets with respect to a constant depth d and rotating around it, resulting in stable training. In Section 4, we define a *trivial baseline* for our approach which assumes a constant depth for all points (depth offsets equals zero, *flat* human skeleton output) and show that our approach can predict meaningful depths offsets.

Discriminator: The discriminator D is defined as a neural network that consumes either the fake 2D pose $\mathbf{p}$ (randomly projected from generated 3D skeleton) or a real 2D pose $\mathbf{r}$ (some projection, via camera or synthetic view, of a real 3D skeleton) and classifies them as either fake (target probability of 0) or real (target probability of 1), respectively.

$$\mathrm{D}_{\theta_D}(\mathbf{u}) \rightarrow [0, 1] \tag{3}$$

where θ_D are parameters of the discriminator learned during training and $\mathbf{u}$ denotes a 2D pose. Note that for any training sample $\mathbf{x}$, we do not require $\mathbf{r}$ to be same as $\mathbf{x}$ or any of its multi-view correspondences. During learning we utilize a standard GAN loss [12] defined as

$$\min_G \max_D V(D, G) = \mathbb{E}(\log(D(\mathbf{r}))) + \mathbb{E}(\log(1 - D(\mathbf{p}))) \tag{4}$$

Priors on 3D skeletons such as the ratio of limb lengths and joint angles are implicitly learned using only random 2D projections.

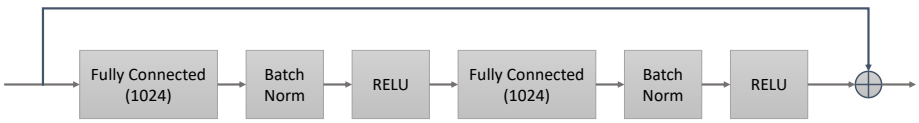

Fig. 3. Residual block used in our generator and discriminator architecture

3.1 Training

For training we normalize the 2D pose landmarks by centering them using the root joint and scaling the pixel coordinates so that the average head-root distance on training data is $1/d$ units in 2D. Although we can fit the entire data in GPU memory, we use a batch size of 32,768. We use the Adam optimizer [21] with a starting learning rate of 0.0002 for both generator and discriminator networks. We varied the batch size between 8,192 and 65,536 in experiments but it did not have any significant effect on the performance. Training time on 8 TitanX GPUs is 0.4 seconds per batch.

Generator Architecture: The generator accepts a 28 dimensional input representing 14 2D joint locations. Inputs are connected to a fully connected layer to expand the dimensionality to 1024 and then fed into subsequent residual blocks. Similar to [26], a residual block is composed of a pair of fully connected layers, each with 1024 neurons followed by batch normalization [16] and `RELU` (see Figure 3). The final output is reduced through a fully connected layer to produce 14 dimensional depth offsets (one for each pose joint). A total of 4 residual blocks are employed in the generator.

Discriminator Architecture: Similar to the generator, the discriminator also takes 28 inputs representing 14 2D joint locations, either from the real 2D pose dataset or the fake 2D pose projected from the hypothesized 3D skeleton. This goes through a fully connected layer of size 1024 to feed the subsequent 3 residual blocks as defined above. Finally, the output of the discriminator is a 2-class softmax layer denoting the probability of the input being real or fake.

Random Rotations: The random projection layer creates a random rotation by sampling an elevation angle ϕ randomly from [0,20] degrees and an azimuth angle θ from [0,360] degrees. These angles were chosen as a heuristic to roughly emulate probable viewpoints that most "in then wild" images would have.

4 Experimental Results

We present quantitative and qualitative results on the widely used Human3.6M [17] for benchmarking. We also show qualitative visualization of reconstructed 3D skeleton from 2D pose landmarks on MPII [3] and Leeds Sports Pose [19] datasets, for which the ground truth 3D data is not available.

4.1 Dataset and Evaluation Metrics

The Human3.6M dataset is one of the largest Human Pose datasets, consisting of 3.6 million 3D human poses. The dataset contains video and 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*.

We found multiple variations of the evaluation protocols in recent literature. We report results on the two most popular protocols. Our **Protocol 1** reports test results only on subject S11 to allow comparison with [7, 45]. **Protocol 2** reports results for both S9 and S11 as adopted by [26, 47, 40, 10, 22]. In both cases, we report the Mean Per Joint Position Error (MPJPE) in millimeters after scaling and rigid alignment to the ground truth skeleton. As discussed, our approach generates 3D skeleton up to a scale factor, since it is impossible to estimate the global scale of a human from a monocular image without additional information. Our results are based on 14-joints per skeleton. We do not train class specific models or leverage any motion information to improve our results. The reported metrics are taken from the respective papers for comparisons.

Similar to previous works [10, 45, 22], we generate synthetic 2D training data by projecting randomly rotated versions of 3D skeletons. These 2D poses are used to augment the 4 camera data already available in Human3.6M. We use additional camera positions to augment data from each 3D skeleton (we use 8 cameras compared to 144 in [45]). The rotation angles for the cameras are sampled randomly in azimuth between 0 to 360 degrees and in elevation between 0 to 20 degrees. We only use data from subjects S1, S5, S6, S7, and S8 for training.

Trivial baseline: We define a trivial baseline with a naive algorithm that predicts a *constant* depth for each 2D pose point. This is equivalent to a generator that outputs constant depth offsets. The MPJPE of such a method is 127.3mm for **Protocol 2** using ground truth 2D points. We achieve much lower error rates in practice, reinforcing the fact that our generator is able to learn realistic 3D poses as expected.

4.2 Quantitative Results: Protocol 1

We first compare our approach to methods that adopt Protocol 1 in their evaluation. Table 1 compares the per class and weighted average MPJPE of our method with recent supervised learning methods [7, 45], using ground truth 2D points for test subject S11. Our results are superior in each category and reduces the previous error by 40% (34.2mm vs. 57.5mm). Table 2 compares with the same methods using 2D points obtained from stacked hourglass(SH) [31] pose detector. We similarly reduce the best reported error by 25% (62.3mm vs. 82.7mm). Our method outperforms these supervised approaches in all activities, except *Walking*.

4.3 Quantitative Results: Protocol 2

Next, we compare against weakly supervised approaches such as [10, 48] that exploit 3D cues indirectly, without requiring direct 2D-3D correspondences. Ta-

Table 1. Comparison of our weakly supervised approach to supervised approaches that adopt **Protocol 1**. Inputs are 2D ground truth pose points

Method	Direct.	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Yasin *et al.* [45]	60.0	54.7	71.6	67.5	63.8	61.9	55.7	73.9
Chen *et al.* [7]	53.3	46.8	58.6	61.2	56.0	58.1	48.9	55.6
Ours	**34.3**	**36.4**	**28.4**	**33.7**	**30.0**	**43.8**	**31.7**	**32.5**

Method	SitDown	Smoke	Photo	Wait	Walk	WalkD	WalkP	Avg.
Yasin *et al.* [45]	110.8	78.9	96.9	67.9	47.5	89.3	53.4	70.5
Chen *et al.* [7]	73.4	60.3	76.1	62.2	35.8	61.9	51.1	57.5
Ours	**48.9**	**32.1**	**43.8**	**36.0**	**25.1**	**34.1**	**30.3**	**34.2**

Table 2. Comparison of our weakly supervised approach to supervised approaches that adopt **Protocol 1**. Inputs are 2D detected pose points. SH denotes stacked hourglass pose detector

Method	Direct.	Discuss	Eat	Greet	Phone	Pose	Purchase	Sit
Yasin *et al.* [45]	88.4	72.5	108.5	110.2	97.1	81.6	107.2	119.0
Chen *et al.* [7]	71.6	66.6	74.7	79.1	70.1	67.6	89.3	90.7
Ours (SH)	**58.4**	**59.4**	**58.7**	**64.5**	**59.0**	**60.9**	**57.0**	**61.6**

Method	SitDown	Smoke	Photo	Wait	Walk	WalkD	WalkP	Avg.
Yasin *et al.* [45]	170.8	108.2	142.5	86.9	92.1	165.7	102.0	108.3
Chen *et al.* [7]	195.6	83.5	93.3	71.2	**55.7**	85.9	62.5	82.7
Ours (SH)	**85.8**	**60.4**	**64.7**	**57.4**	63.0	**65.5**	**62.1**	**62.3**

Table 3. Comparison of our approach to other weakly supervised approaches that adopt **Protocol 2**. Inputs are 2D ground truth pose points. Results marked as * are taken from [10]

Method	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Purchase
3DInterpreter [43]*	56.3	77.5	96.2	71.6	96.3	106.7	59.1	109.2
Monocap [48]*	78.0	78.9	88.1	93.9	102.1	115.7	71.0	90.6
AIGN [10]	53.7	71.5	82.3	58.6	86.9	98.4	57.6	104.2
Ours	**33.5**	**39.3**	**32.9**	**37.0**	**35.8**	**42.7**	**39.0**	**38.2**

Method	Sit	SitDown	Smoke	Wait	Walk	WalkD	WalkP	Avg.
3DInterpreter [43]*	111.9	111.9	124.2	93.3	58.0	-	-	88.6
Monocap [48]*	121.0	118.2	102.5	82.6	75.62	-	-	92.3
AIGN [10]	100.0	112.5	83.3	68.9	57.0	-	-	79.0
Ours	**42.1**	**52.3**	**36.9**	**39.4**	**36.8**	**33.2**	**34.9**	**38.2**

ble 3 compares the MPJPE for the previous weakly supervised approaches using **Protocol 2** on ground truth 2D pose inputs. Our approach reduces the error

Table 4. Comparison of our approach to other weakly supervised approaches that adopt **Protocol 2**. Inputs are 2D detected pose points. SH denotes stacked hourglass. Results marked as * are taken from [10]

Method	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Purchase
3DInterpreter [43]*	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7
AIGN [10]	77.6	91.4	89.9	88	107.3	110.1	75.9	107.5
Ours (SH)	**60.2**	**60.7**	**59.2**	**65.1**	**65.5**	**63.8**	**59.4**	**59.4**

Method	Sit	SitDown	Smoke	Wait	Walk	WalkD	WalkP	Avg.
3DInterpreter [43]*	127.4	139.0	103.4	91.4	79.1	-	-	98.4
AIGN [10]	124.2	137.8	102.2	90.3	78.6	-	-	97.2
Ours (SH)	**69.1**	**88.0**	**64.8**	**60.8**	**64.9**	**63.9**	**65.2**	**64.6**

reported in [10] by more than 50% (38.2mm vs. 79.0mm). A similar comparison is shown in Table 4 using 2D key points detected using the stacked hourglass [31] pose estimator. Our approach outperforms other methods in all activity classes and reduces the previously reported error by 33% (64.6mm vs. 97.2mm).

It is well known that supervised approaches broadly perform better than weakly supervised approaches in classification and regression tasks. For human 3D pose estimation, we do not expect our method to outperform the state of the art supervised approach [26]. However, our results are better than several previous published works that use 3D supervision as shown in Table 5. We have demonstrated the effectiveness of a relatively simple adversarial training framework using only 2D pose landmarks as input.

While the focus of this work is on weakly supervised learning from 2D poses alone, we are very encouraged by the fact that our results are competitive with the state of the art supervised approaches. In fact, our approach comes to within 1.1mm of the error reported by [26] on the ground truth 2D input as shown in Table 6. We also experimented with a naïve ensemble algorithm where we combined 11 of our top performing models on the validation data and averaged the 3D skeleton for each input. This simple algorithm reduced the error to 36.3mm, surpassing the state-of-the-art results of 37.1mm (Table 6).

4.4 Qualitative Results

Figure 4 shows a few 3D pose reconstruction results on Human3.6M using our approach. The ground truth 3D skeleton is shown in gray. We see that our approach can successfully recover the 3D pose. Figure 5 shows some failure cases of our approach. Our typical failures are due to odd or challenging poses containing severe occlusions or plausible alternate hypothesis such as mirror flips in the direction of viewing. Since the training was performed on images containing all 14 joints, we are currently unable to lift 2D poses with fewer joints to 3D skeletons.

Table 5. Comparison of our approach to other supervised methods on Human3.6M under **Protocol 2** using detected 2D keypoints. The results of all approaches are obtained from [26]. Our approach outperforms most supervised methods that use explicit 2D-3D correspondences

Method	Direct.	Discuss	Eat	Greet	Phone	Photo	Pose	Purchase
Akhter & Black [1]	199.2	177.6	161.8	197.8	176.2	186.5	195.4	167.3
Ramakrishna *et al.* [36]	137.4	149.3	141.6	154.3	157.7	158.9	141.8	158.1
Zhou *et al.* (2016)[47]	99.7	95.8	87.9	116.8	108.3	107.3	93.5	95.3
Bogo *et al.* [4]	62.0	60.2	67.8	76.5	92.1	77.0	73.0	75.3
Moreno-Noguer [30]	66.1	61.7	84.5	73.7	65.2	67.2	60.9	67.3
Martinez *et al.* [26]	44.8	52.0	44.4	50.5	61.7	59.4	45.1	41.9
Ours (Weakly Supervised)	*60.2*	*60.7*	*59.2*	*65.1*	*65.5*	*63.8*	*59.4*	*59.4*

Method	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkP	Avg.
Akhter & Black [1]	160.7	173.7	177.8	181.9	176.2	198.6	192.7	181.1
Ramakrishna *et al.* [36]	168.6	175.6	160.4	161.7	150.0	174.8	150.2	157.3
Zhou *et al.* (2016) [47]	109.1	137.5	106.0	102.2	106.5	110.4	115.2	106.7
Bogo *et al.* [4]	100.3	137.3	83.4	77.3	86.8	79.7	87.7	82.3
Moreno-Noguer [30]	103.5	74.6	92.6	69.6	71.5	78.0	73.2	74.0
Martinez *et al.* [26]	66.3	77.6	54.0	58.8	49.0	35.9	40.7	52.1
Ours (Weakly supervised)	*69.1*	*88.0*	*64.8*	*60.8*	*64.9*	*63.9*	*65.2*	*64.6*

Table 6. Comparison of our results to the state of the art fully supervised approaches under Protocol 2 using ground truth 2D inputs. Our model has error within 1.1mm of the best supervised approach, and surpasses it with a naïve ensemble approach

Moreno-Noguer [30]	Martinez *et al.* [26]	Ours (Weakly supervised) (Single Model)	Ours (Weakly supervised) (Ensemble)
62.2	<u>37.1</u>	38.2	**36.3**

To test the generalization performance of our method on images in the wild, we applied our method to MPII [3] and the Leeds Sports Pose (LSP) [19] datasets. MPII consists of images from short Youtube videos and has been used as a standard dataset for 2D human pose estimation. Similarly LSP dataset contains images from Flickr containing people performing sport activities. Figures 6 and 7 show some representative examples from these datasets containing humans in natural and difficult poses. Despite the change in domain, our weakly supervised method successfully recovers 3D poses. Note, our model is not trained on 2D poses from MPII or LSP datasets. This demonstrates the ability of our method to generalize over characteristics such as object distance, camera parameters, and unseen poses.

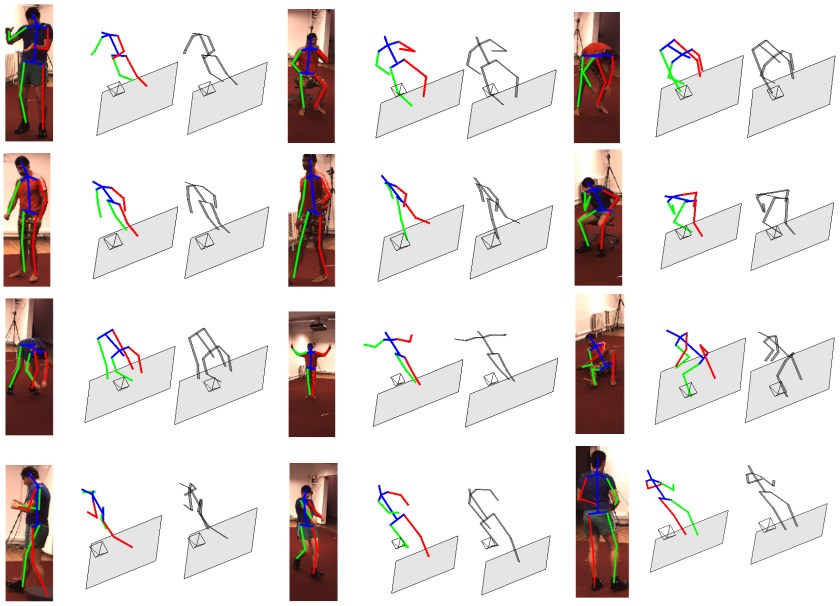

Fig. 4. Examples of 3D pose reconstruction on Human3.6M dataset. For each image we overlay 2D pose points, followed by the predicted 3D skeleton in color. Corresponding ground truth 3D shown in gray

4.5 Discussion

As a general observation, we noticed that the results for the *SitDown* class were the worst across the board for all methods on Human3.6M dataset. In addition to the obvious explanation of fewer available examples in this class, sit down poses lead to significant occlusion of the MoCap markers on legs or ankles (see for example Figure 5). This phenomenon leads to some of the high errors for this class.

Overall our qualitative and quantitative results have substantiated the effectiveness of using adversarial training paradigm for learning 3D priors in a weakly supervised way. Not only do we outperform the majority of similar 2D-3D lifting methods in benchmarks, we have also shown robust performance on "images in the wild" datasets. Even though we do not leverage any temporal information when available, we found the results from our approach to be stable when applied on per frame basis to video sequences.

In summary, we believe that we have pushed the state of art in 3D pose estimation using weakly supervised learning. This paves way for new research directions that can extend this work by combining supervised, weakly-supervised, and unsupervised frameworks (*i.e.*, semi-supervised) for lifting 2D poses to 3D skeletons.

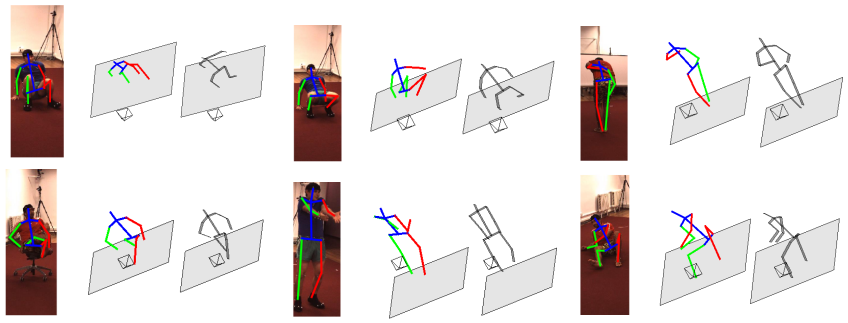

Fig. 5. Some failure cases for our approach on Human3.6m dataset. Ground truth 3D skeleton is shown in gray

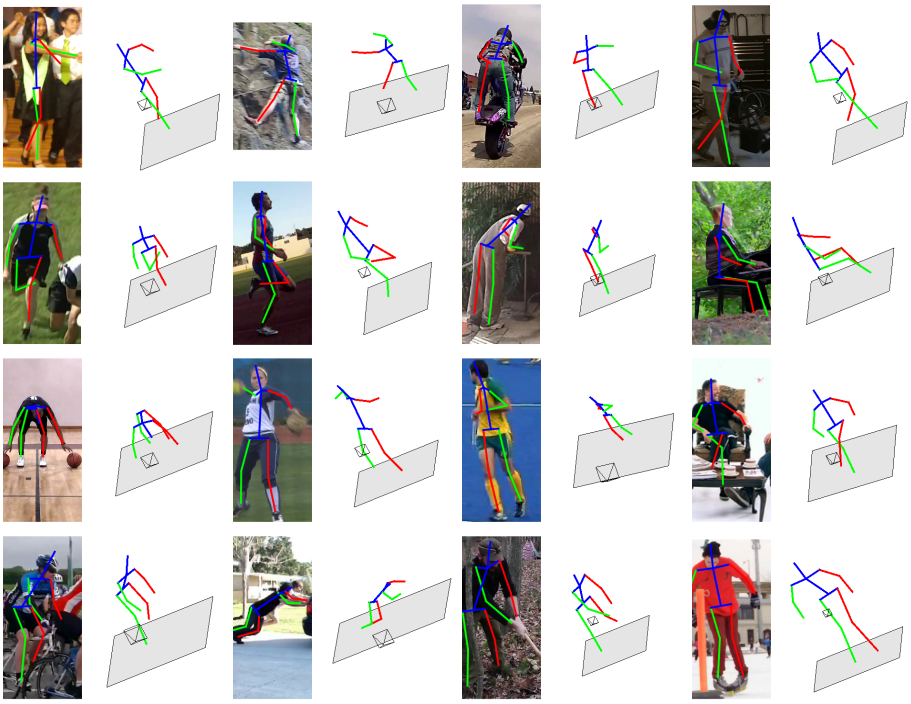

Fig. 6. Examples of 3D pose reconstruction on MPII dataset. For each image we overlay 2D pose points, followed by the predicted 3D skeleton in color. The dataset does not contain ground truth 3D skeletons

5 Conclusions

While most of the recent progress in deep learning is fueled by labeled data, it is difficult to obtain high quality annotations for most computer vision problems.

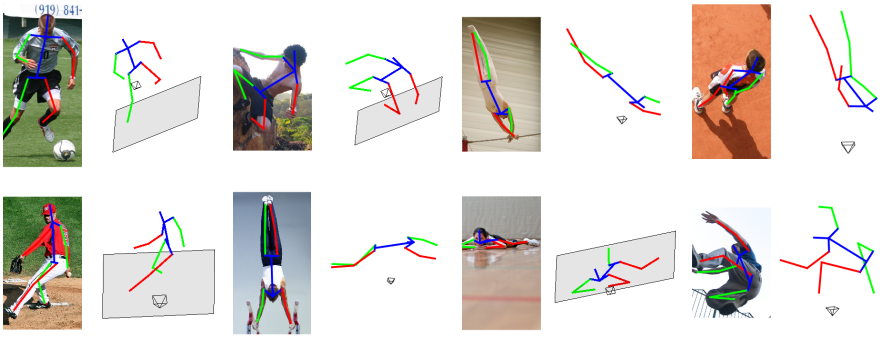

Fig. 7. Examples of 3D pose reconstruction on Leeds dataset. For each image we overlay 2D pose points, followed by the predicted 3D skeleton in color. The dataset does not contain ground truth 3D skeletons

For 3D human pose estimation, acquiring 3D MoCap data remains an expensive and challenging endeavor. Our paper demonstrates that an effective prior on 3D structure and pose can be learned from random 2D projections using an adversarial training framework.

We believe that our paper presents a unique insight into learning 3D priors via projective geometry and opens up numerous interesting applications, beyond human pose estimation. Applications such as sparse 3D reconstruction for indoor scenes as well as outdoor navigation typically requires a 3D sensor or multi-view images with a structure from motion pipeline. We envision that our approach can be applied for learning sparse 3D reconstructions from edges or keypoints extracted from a single image, using a collection of 2D images. Since the inference only requires running a small neural network, our approach can also be used for interactive graphics for rendering 3D models from line drawings, where the 3D prior is learned from a collection of such line drawings. We anticipate that our paper will spark further interest in applications combining learning techniques with projective geometry.

Finally, inspired by the results presented in this paper, we expect future work to explore the use of solely images "in the wild" for training the system. This could be achieved through an end to end, image to 3D adversarial pipeline or the use of a large annotated 2D pose dataset. Additional research could include analysis of difficult poses not generally seen (hand-stands, gymnast or yoga poses) as well as the use of additional datasets such as the MPI-INF-3DHP [27]. Improving the robustness of the system by incorporating compensation for noisy or imperfect 2D pose inputs or predicting a distribution of joint locations is another future research opportunity. There is also promise for our method to improve the state of the art by combining with fully supervised approaches.

References

1. Akhter, I., Black, M.J.: Pose-conditioned joint angle limits for 3d human pose reconstruction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1446–1455 (2015)
2. Amin, S., Andriluka, M., Rohrbach, M., Schiele, B.: Multi-view pictorial structures for 3d human pose estimation. In: BMVC (2013)
3. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: CVPR (2014)
4. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision. pp. 561–578. Springer (2016)
5. Brau, E., Jiang, H.: 3d human pose estimation via deep learning from 2d annotations. In: Fourth International Conference on 3D Vision (2016)
6. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR (2017)
7. Chen, C.H., Ramanan, D.: 3d human pose estimation = 2d pose estimation + matching. In: CVPR (2017)
8. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
9. Fang, H.S., Xu, Y., Wang, W., Liu, X., Zhu, S.C.: Learning pose grammar to encode human body configuration for 3d pose estimation. In: AAAI Conference on Artificial Intelligence (2018)
10. Fish Tung, H.Y., Harley, A.W., Seto, W., Fragkiadaki, K.: 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)
11. Forsyth, D.A., Arikan, O., Ikemoto, L.: Computational Studies of Human Motion: Tracking and Motion Synthesis. Now Publishers Inc (2006)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S.: Generative adversarial nets. In: NIPS. pp. 2672–2680 (2014)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Computer Vision (ICCV), 2017 IEEE International Conference on. pp. 2980–2988. IEEE (2017)
14. Hofmann, M., Gavrila, D.M.: Multi-view 3d human pose estimation in complex environment. IJCV (2012)
15. Hogg, D.: Model-based vision: a program to see a walking person. Image and Vision computing (1983)
16. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML. pp. 448–456 (2015)
17. Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C.: 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)
18. Jiang, H.: 3d human pose reconstruction using millions of exemplars. In: Pattern Recognition (ICPR), 2010 20th International Conference on (2010)
19. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation (2010)
20. Kanazawa, A., Black, M., Jacobs, D., Malik, J.: End-to-end recovery of human shape and pose. In: TBD (2018)

21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. ICLR (2015), http://arxiv.org/abs/1412.6980
22. Li, S., Chan, A.B.: 3d human pose estimation from monocular images with deep convolutional neural network. In: ACCV (2014)
23. Li, S., Zhang, W., Chan, A.B.: Maximum-margin structured learning with deep networks for 3d human pose estimation. In: IEEE International Conference on Computer Vision (ICCV) (December 2015)
24. Lin, M., Lin, L., Liang, X., Wang, K., Cheng, H.: Recurrent 3d pose sequence machines. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
25. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM Transactions on Graphics (TOG) **34**(6), 248 (2015)
26. Martinez, J., Hossain, R., Romero, J., Little, J.: A simple yet effective baseline for 3d human pose estimation. In: ICCV (2017)
27. Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., Theobalt, C.: Monocular 3d human pose estimation in the wild using improved cnn supervision. In: 3D Vision (3DV), 2017 International Conference on. pp. 506–516. IEEE (2017)
28. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.P., Xu, W., Casas, D., Theobalt, C.: Vnect: Real-time 3d human pose estimation with a single rgb camera. ACM Transactions on Graphics (TOG) **36**(4), 44 (2017)
29. Moeslund, T.B., Granum, E.: A survey of computer vision-based human motion capture. Computer Vision and Image Understanding (2001)
30. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
31. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: European Conference on Computer Vision. pp. 483–499. Springer (2016)
32. O'Rourke, J., Badler, N.I.: Model-based image analysis of human motion using constraint propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence (1980)
33. Park, S., Hwang, J., Kwak, N.: 3d human pose estimation using convolutional neural networks with 2d pose information. In: European Conference on Computer Vision. pp. 156–169. Springer (2016)
34. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR (July 2017)
35. Rafi, U., Gall, J., Leibe, B.: A semantic occlusion model for human pose estimation from a single depth image. In: Proceedings of the IEEE Conference on CVPR Workshops (2015)
36. Ramakrishna, V., Kanade, T., Sheikh, Y.: Reconstructing 3d human pose from 2d image landmarks. In: European Conference on Computer Vision. pp. 573–586. Springer (2012)
37. Rogez, G., Weinzaepfel, P., Schmid, C.: Lcr-net: Localization-classification-regression for human pose. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
38. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M., Moore, R.: Real-time human pose recognition in parts from single depth images. Communications of the ACM (2013)

39. Tekin, B., Marquez-Neila, P., Salzmann, M., Fua, P.: Learning to fuse 2d and 3d image cues for monocular body pose estimation. In: IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
40. Tekin, B., Rozantsev, A., Lepetit, V., Fua, P.: Direct prediction of 3d body poses from motion compensated sequences. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 991–1000 (2016)
41. Tome, D., Russell, C., Agapito, L.: Lifting from the deep: Convolutional 3d pose estimation from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017)
42. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: CVPR. pp. 4724–4732 (2016)
43. Wu, J., Xue, T., Lim, J.J., Tian, Y., Tenenbaum, J.B., Torralba, A., Freeman, W.T.: Single image 3d interpreter network. In: ECCV. pp. 365–382 (2016)
44. Xiaohan Nie, B., Wei, P., Zhu, S.C.: Monocular 3d human pose estimation by predicting depth on joints. In: IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
45. Yasin, H., Iqbal, U., Kruger, B., Weber, A., Gall, J.: A dual-source approach for 3d pose estimation from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016)
46. Yub Jung, H., Lee, S., Seok Heo, Y., Dong Yun, I.: Random tree walk toward instantaneous 3d human pose estimation. In: CVPR (2015)
47. Zhou, X., Zhu, M., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Sparseness meets deepness: 3d human pose estimation from monocular video. In: CVPR (2016)
48. Zhou, X., Zhu, M., Pavlakos, G., Leonardos, S., Derpanis, K.G., Daniilidis, K.: Monocap: Monocular human motion capture using a cnn coupled with a geometric prior. In: TBD (2017)
49. Zhou, X., Huang, Q., Sun, X., Xue, X., Wei, Y.: Towards 3d human pose estimation in the wild: a weakly-supervised approach. In: IEEE International Conference on Computer Vision (2017)