7th International Conference on Knowledge and Innovation in Engineering, Science and Technology
15 - 17 December, 2020 Berlin, Germany

Multi-view Generative Networks for 3D Pose Estimation

Motaz Sabri

[1]RidgeI, Tokyo, Chiyoda 161 Japan

Abstract.

Monocular 3D human pose estimation in the wild is still a challenging task due to the scarcity of annotated yet unconstrained training data for accurate 3D poses. In this paper, we tackle this issue by proposing a weakly-supervised approach that learns to estimate 3D poses from unlabeled multi-view generated data from a single RGB image without relying on 3D annotations. Since the generation of multi-view data from a single image is prone to degenerated solutions, we utilize a GAN based approach to create multi-view pose representations that are authentic. The added constraints on the latent distribution simplify the learning of a shared latent space between the depth map and the pose. It also improves the approach generalization and exploitation of unlabeled depth maps. We evaluate our approach on three challenging datasets (Human3.6M, MPII-INF-3DHP and Leeds SportsPose) where it achieves state of the art performance among semi and weakly-supervised methods.

Keywords: Panoptic reconstruction, View generation, 3D reconstruction, inpainting, VAE.

1. Introduction

Monocular images have been used for human pose estimation actively in many recent computer vision researches with many applications such as security, medicineand human-computer interaction. There are numerous approaches to handle generating 3D human poses from monocular images (Park et al., 2016, Mehta et al., 2017a,b, Pavlakos et al., 2017a, MorenoNoguer. et al., 2017, Martinez et al., 2017, Luo et al., 2018, Rayat Imtiaz Hossain et al., 2018). The supervised learning approaches are taking the lead in this field due to the availability of a large corpus of depth images annotated with body joints. What's common among most of these approaches beside their accurate results on similar data is that they lack the ability to generalize to unknown movements and view angles. Weakly-supervised learning provides an alternative method for learning robust geometry representation without extensive precise 3D annotation. Many approaches (Wandt et al., 2019, Habibie et al., 2019, Chen et al., 2019, Kocabas,2019, Moon et al., 2019, Pavllo et al., 2019) leverage knowledge transformation to increase their robustness by training 3D annotations with abundant 2D annotations. These methods face challenges in domain shift between training poses and in-the-wild poses.

Overcoming the generalization challenge requires a great deal of annotation, which is tedious and error prone. In this work, we use synthesized multi-view images generated from original RGB image to provide weak supervision and estimate the 3D pose of the body. We use variational autoencoder (VAE) (Diederik et al., 2014) generative model to create the views and generative adversarial network (GAN) (Goodfellow et al., 2014) to capture the latent spaces of body poses and corresponding images for estimating 3D body pose. We also propose a mapping between the latent body pose space and latent joint-mapping space.This mapping is rewarding for learning, since a key point that is sampled in any of the two latent spaces can be expressed either as a 3D pose via the VAE's decoder or as a joint-mapping through GAN.

Our experiments demonstrate the effectiveness of our weakly supervised multi-view training strategy on Human3.6M (Ionescu et al., 2014), MPII-INF-3DHP (MPII) (Mehta et al., 2017a), and Leeds Sports Pose (LSP) (Johnson,2010) datasets. Besides drastically reducing the need for annotated data, our approach also increases the robustness against viewpoint and scaling.

The remainder of this paper is organized as follows. Related works are introduced in Section 2 and details of the proposed method are provided in Section3. Experimental results are presented in Sections 4, followed by the conclusionin Section 5.

2. Related Work

Both supervised and semi-supervised approaches have been used to achieve a high-quality 3D pose estimation. Supervised methods rely on deep learning architectures that utilize very large datasets for training (Martinez et al., 2017, Mehta et al., 2017b, Pavlakos et al., 2017a, Popa et al., 2017, Tome et al., 2018, Sun et al., 2017, Tekin et al., 2017, Zhou et al., 2017b, Pavlakos et al., 2017b). As those supervised approaches are heavily dependent on the training data, they struggle to generalize outside the poses and motions of training set. Researchers have invested in adding more content to the training data through more annotations (Mehta et al., 2017a, Joo, 2015, Pavlakos et al., 2017b) and data augmentation (Ionescu et al., 2015, Rogez et al., 2016). However, the limited diversity of appearance and motion that current tools provide, along with their imperfect verisimilitude, limits both generality and accuracy of networks trained using only those images.

Due to the previous factors, weak supervision is considered a promising alternative in which network is trained without 2D to 3D supervision (Wandt et al., 2019, Habibie et al., 2019, Chen et al., 2019, Kocabas et al., 2019, Moon et al., 2019, Pavllo et al., 2019). Researchers utilized weak supervision to let their models acquire knowledge through different perspectives. (Kocabas et al., 2019) propose Epipolar geometry to self-supervise a 3D pose estimator. (Wandt et al., 2019) pair a weakly supervised model with a reprojection loss for training and use adversarial losses to distinguish between plausible and implausible poses. (Kanazawa et al., 2018) use a single RGB image to create a 3D mesh that is parameterized in terms of 3D joint angles and a low dimensional linear shape space. (Pavllo et al., 2019) utilize temporal information with dilated convolutions over 2D key point trajectories to estimate 3D poses in videos. They also propose a semi supervised training method to improve performance when labeled data is scarce. The robustness of 3D pose estimation has increased through the above proposals however its possible to improve generalization against viewpoints and scale changes.

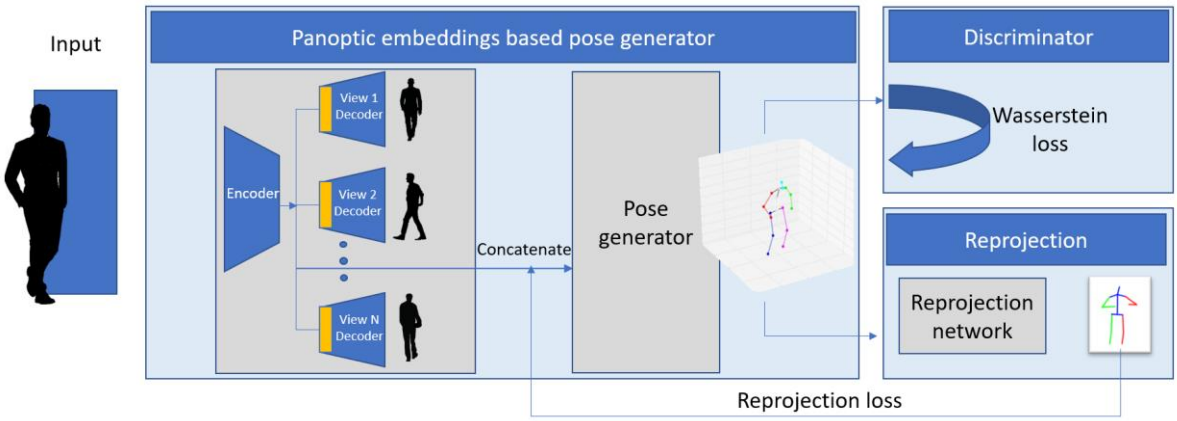
In this work, we propose a weakly-supervised method that is trained with generated multi-view data from a single RGB image. (Tome et al., 2018) trained using motion capture data to generate probabilistic 3D pose estimator. Their model is integrated into a multi-staged 2D pose estimation process to enhance 2D and 3D pose predictions iteratively. (Novotny et al., 2019) factorize the problem of pose estimation across multi viewpoint and shape parameters. They used canonicalization loss to provide inductive bias. (Wan et al., 2017) used two generative models to allow learning a mapping between the two latent spaces. They exploit the generalization properties of the GAN and pose constraints implicitly learned by the VAE to improve discriminative pose estimation.

Both (Kocabas et al., 2019, Pavlakos et al., 2017b) use unlabeled generated data for training. The approach of (Pavlakos et al., 2017b), however, attempts to configure camera parameters in unconstrained environments. (Kocabas et al., 2019) uses Epipolar estimates 2D poses and reconstructs corresponding 3D pose. Their geometry consider 3D poses to remain fixed throughout the training. As a result, the errors in 3D reconstruction directly propagate to the trained models causing 2D pose estimation to fail easily. In this work, we propose a learning procedure which can handle poses variety in the captured data. We train using two generative models unconstrained with view constrains. This is optimized for 2D and 3D poses. Our approach does not require labeled data to improve 2D predictions. It outperforms methods trained for weakly supervised learning by large margin.

3. Proposed method

We propose to regress 3D poses from 2D observations by learning a mapping from the input distribution to the 3D pose distribution. To achieve such learning, we introduce an intermediate distribution by generating N views of an observation using VAE and GAN. In standard GAN (Goodfellow et al., 2014) training, the input is sampled from a Gaussian or uniform distribution. In our method, the generator input is obtained from the latent space of the VAE while generating N sided panoptic information from a single RGB image. Giving our method the name panoptic embeddings-based pose estimator (PEPE). This choice of generator generates 3D poses with some incorrect 3D reconstructions of the input 2D observations. To ensure a correct match between 2D and 3D poses we adopt the Wasserstein loss for the GAN (Wandt et al., 2019, Arjovsky et al., 2017). Figure 1 shows the proposed network consisting of three parts trained alternatively: generator network, discriminator used in the adversarial training and finally reprojection part that maps from a distribution of 2D poses to a distribution of 3D poses.

Figure 1: The proposed architecture consists of three parts: Panoptic embeddings based pose generator: 3D pose estimator that uses features extracted from a multi view VAE, Discriminator with Wasserstein loss that scores the realness of a given image instead of the probability of the image realness. Lastly the reprojection network guided by (Wandt, 2019).



3.1 Generator

The generator models a prior distribution on body pose configurations using inception (Szegedy. et al., 2015) VAE. Its structure allows learning the mapping from high dimensional body poses to a low-dimensional representation through its decoders. Let g represent some generated observation. We want to estimate a prior $P(g)$ by modeling the generation process of g by sampling some h from an arbitrary low-dimensional distribution $P(h)$ as

$$P(g) = \int_h P(g|h)\, P(h) \partial h \tag{1}$$

Fitting $P(g)$ directly usually involves expensive inference. Therefore, we approximate $P(g)$ using variational autoencoder (VAE) guided by a generative adversarial network (GAN). We provide a brief mathematical representation for the usage of VAE and GAN and how we use them to model the prior of body poses and depth mapping for the poses. Figure 2 illustrates details of the VAE and the pose generator. We denote the generated depth map as y from 2D input image x. The VAE generates N outputs representing N camera views of the person in x. Therefore, we have N decoders and a single encoder. $\overline{x_n}$ refers to the reconstructed pose parameter from the n^{th} decoder. y refers to the synthesized depth map from the GAN generator. h_{x_n} and h_{y_n} indicate the n^{th} latent pose of view and depth map respectively.

Our VAE regularizes single encoder and N decoders to estimate latent variable posterior as:

$$Enc(x) = \sum_{n=1}^{N} q(h_{x_n}|x) \tag{2}$$

$$Dec(h_{x_n}) = P(x|h_{x_n}) \tag{3}$$

The latent pose $\sum_{n=1}^{N} h_{x_n}$ is regulated by introducing a prior over the latent distribution on $P(h_{x_n})$ and reconstructing $\overline{x_n}$ to make it correspond to the view of the given 2D image x. Commonly, a Gaussian prior is integrated into the loss as the Kullback-Leibler divergence DKL between the encoded distribution $q(h_{x_n}|x)$ and the prior $p(h_{x_n})$. The VAE loss is given by reconstruction errors summation of N decoders and latent prior $\mathcal{L}_{vae} = \sum_{n=1}^{N} \mathcal{L}_n^{pose} + \mathcal{L}_p$

where $\mathcal{L}_n^{pose}$ is the reconstruction loss of the n^{th} view and its given as:

$$\mathcal{L}_n^{pose} = -\mathbb{E}_{q(h_{x_n}|x)}[\log P(x|h_{x_n})] \tag{4}$$

and $\mathcal{L}_p$ is prior loss and its given as:

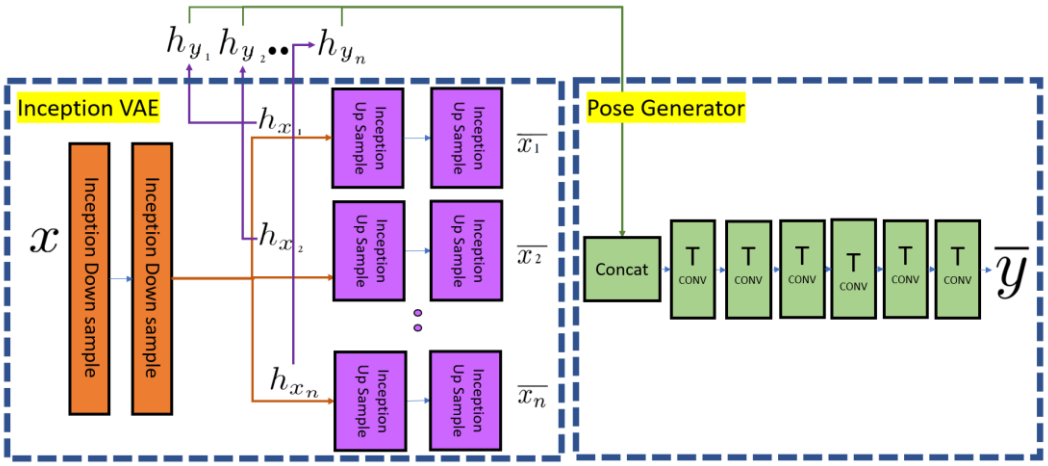
$$\mathcal{L}_p = \sum_{n=1}^{N} Dkl\left((q(h_{x_n}|x)||p(h_{x_n})\right) \tag{5}$$

We want to reconstruct the depth map using VAE extracted latent variables through GAN. However, GAN alone can't perform this estimation for latent variable posterior. Therefore, we must impose learning a mapping from H_x latent space (which is formed by concatenating all the latent spaces h_{x_n}) to H_y . We use the latent space parameter of the body pose as the reference space to learn a mapping to the depth map latent space that is H_y through MAP (H_x).

Having the corresponding pairs x and y we can train using observed depth images y as teacher signal and with synthesized images $\overline{y} = Gen(MAP(H_x))$ that are projected to H_x and then mapped to H_y. We introduce an intermediate loss $\mathcal{L}_r$, based on reconstruction error of the rendered depth map given a latent input H_x which is mapped to the GAN latent space as follows:

$$\mathcal{L}_r = \max(||y_n - Gen(MAP(H_x))||^2, \xi \tag{6}$$

Figure 2: The panoptic embeddings-based pose generator (referenced in figure 1) is shown. The VAE component and the generator create authentic $\overline{y}$ to fool the discriminator.



Where ς is the clipping threshold and $MAP(.)$ is a single fully connected neuron with tanh activation marked in purple arrow in figure 2. We train this network using a clipped mean squared error loss function. This ensures robustness to depth estimation noise as used in (E. Simo-Serra et al., 2012). Since the depth map is normalized to$[1, -1]$, we set ξ = 1. Since our generative model can learn low-dimensional representations, we are able to generate realistic samples with a very small set of labeled (x,y) pairs. After learning the mapping, we can project any point in the latent pose space into both a body pose space or into a corresponding depth map space. We can regard the two compatible latent spaces as a common shared latent pose.

The composite function $Gen(MAP(.))$ generates the depth latent space. Its input is the mapped latent space marked in by the green arrow in 2. Note that $MAP(.)$ is implicitly learning a mapping from a normal distribution H_x to H_y. Therefore, any random noise sampled from the standard normal distribution can be mapped either to a body pose or a corresponding depth map. The $Gen(MAP(.))$ function is implemented as six transposed convolutional layers with dilation factor of two in order to build $\overline{y}$. The constraint on the latent distribution simplify the learning of a shared latent space between the depth map and the pose. The discriminator tries to distinguish between real data samples y and synthetic samples $\overline{y}$created by the generator.

3.2 Discriminator

We chose to design the discriminator to be like the pose regression network. We train it using the Wasserstein loss function (Arjovsky et al., 2017). We incorporate kinematic chain space (KCS) (Wandt and Rosenhahn. et al., 2018) to handle joint angle dynamics, kinematic chains, and symmetry. KCS derives a constraint that assumes that bone lengths are constant. This formulates an easy to solve nuclear norm optimization problem. It allows for better scenes reconstruction without depending on predefined body measures. The KCS layer with a subsequent fully connected network forms another branch of the discriminator network. These two branches are merged before the output layer. Consecutively, the GAN loss is:

$$\mathcal{L}_{GAN} = \log\big(\mathrm{Disc(y)}\big) + \log(1 - Disc(Gen(H_y))) \quad (7)$$

where Disc(y) is discriminator output and $Gen(H_y)$ is the generated pose. Generator aims to reduce the loss while discriminator tries to confuse the generator by maximizing the loss.

3.3 Reprojection

The reprojection layer takes as input the synthetic poses $\bar{y}$ created by the generator and remaps it into 2D coordinate space. This follows method of (Wandt et al., 2019). It is given as $w' = y$ where w' denotes 2D reprojection. The reprojection loss function $\mathcal{L}_j(.)$ can be defined as:

$$\mathcal{L}_j(\bar{y}) = ||\, w \; - \bar{y}||_F \tag{8}$$

Where w is the observed 2D pose matrix, which has the same structure as w'. The $||\,.||_F$ denotes the Frobenius norm. The reprojection layer is a single layer without trainable parameters. If a joint is not detected, then its corresponding columns in w and y are set to zero. This marginalizes their influence on the value of the reprojection loss. The missing joints will then be hallucinated by the pose generator network according to the discriminator.

4. Experiments

4.1 Training

We trained our model using a fine-grained subset of the Panoptic studio (Joo, 2015) dataset. The dataset contains 480 synchronized video streams of multiple people engaged in social activities and contains anatomical landmarks labels of individuals in the space. All cameras heights are fixed, and the subject movements happen at defined zones. This offers consistency for generative model to learn salient characteristics robustly while marginalizing the background noise. Hence, we choose this data over the others for training. Our fine-grained streams are from selected 8 cameras that are 2.2-meter-high in the panoptic sphere for none overlapping subjects performing random activities. This allowed us to acquire octuplets samples. At a single time, a random item of the octuplets items is considered an input while the other seven (plus the selected input) forms the VAE output. Per sample, the random selection of the input happens 4 times. The above process yields 40,000 samples in which 35,000 samples were used for training. We have not used other datasets for training as our selection. Our selected samples have a variety of person sizes which gave the 2D joint generator an arbitrary scale. To remove this effect, we divide each generated 2D pose vector by its standard deviation. As a result, the possible 2D pose values are constrained. This allows our method to perform domain transfer of 2D poses more easily.

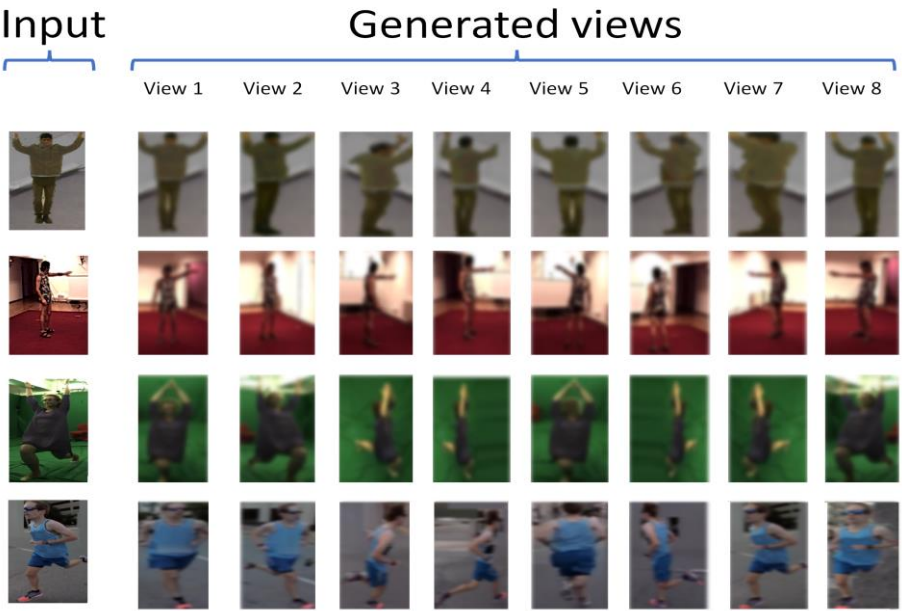
The generator, discriminator and reprojection are trained alternatively. Training alternates between minimizing $\mathcal{L}_{GAN}$ parameters of the generator then maximizing $\mathcal{L}_{GAN}$ w.r.t. parameters of the discriminator and finally minimizes $\mathcal{L}_j$ to remap the generated 3D poses into the 2D space for the generator to start learning from another batch. We stabilize training on every hidden layer by batch normalization. We use the Adam (Kingma et al., 2015) algorithm to optimize our network.

We inject random Gaussian noise with 0.05 standard deviation to the latent variable after VAE encoder during training to increase robustness. We set the learning rate as 0.001 and train the complete network for 5000 epochs with exponential decay every 25 epochs.

4.2 Evaluation

After training our model with Panoptic studio dataset (Joo, 2015), we evaluate its performance on the following three datasets: Human3.6M (Ionescu et al., 2014), MPII (Mehta et al., 2017a), and LSP (Johnson et al., 2010). Human3.6M is the largest benchmark dataset containing images temporally aligned to 2D and 3D correspondences. All the data in those datasets are considered unseen as we use Panoptic studio only for training in all experiments. We evaluate our training quantitatively on Human3.6M and MPII data. For qualitative results containing unusual poses and camera angles we evaluate on LSP. As multi-view generation is an intermediate step and its effect cannot be seen directly through results. Figure 3 shows the PEPE reconstruction of 8 views holding useful structural information from a single RGB image that is used for pose estimation. inputs are unseen data with a variety of poses and projections.

Figure 3: The PEPE can reconstruct 8 views that correspond to the input image. The first images row is from Panoptic studio (testing data). The second images row is from Human3.6M. Images from the third and fourth rows are from the MPII and LSP datasets respectively. None-frontal views are also reconstructed.



Quantitative Evaluation

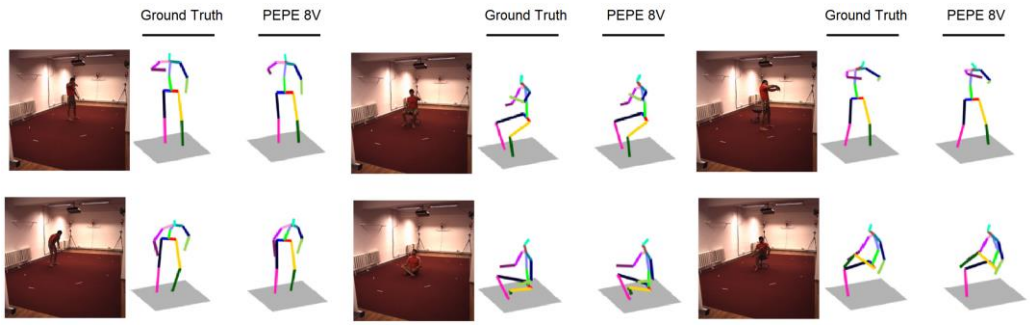
In the literature many of the evaluating protocols calculate the mean per joint positioning error (MPJPE) and Percentage of Correct Key points (PCK) between the reconstructed and the ground truth joint coordinates. The MPJPE calculates per joint position error as the Euclidean distance between ground truth and prediction for a joint. The PCK is evaluated as the percentage of trials where the Euclidean pixel distance between the actual and predicted joint location is below the desired threshold.

The Human3.6M dataset is commonly evaluated using the MPJPE measure while the MPII and LSP datasets are evaluated using the PCK. We adopted the same conventions in our analysis. In our evaluation we use Protocol-1 where no pose alignment happens. Table 1 shows our results on the Human3.6M dataset. We observe generating more views improves the pose estimation. Our method outperforms supervised methods in many cases (Luo et al., 2018, Pavlakos et al., 2017a, Zhou et al., 2017a, Martinez et al., 2017) and unsupervised methods (wandt et al., 2019, Wu et al., 2016, Tung et al., 2017) in most of poses.

Table 1: *MPJPE values for pose estimation of the Human3.6M dataset against state-of-the-art methods. In this comparison we follow Protocol-1(no rigid alignment). Scores are taken from the referenced papers. The rows PEPE-(nV) show our method when we use n decoders to create n views.*

Method	**Direct**	**Disc**	**Eat**	**Greet**	**Phone**	**Photo**	**Pose**	**Purch**
(Luo 2018)	68.4	77.3	70.2	71.4	75.1	86.5	69	76.7
(Pavlakos 2017)	67.4	71.9	66.7	69.1	72	77	65	68.3
(Zhou 2017)	54.8	60.7	58.2	71.4	62	65.5	53.8	55.6
(Martinez 2017	53.3	60.8	62.9	**62.7**	86.4	82.4	57.8	58.7
(Wandt 2019)	77.5	85.2	82.7	93.8	93.9	101	82.9	102.6
(Wu 2016)	78.6	90.8	92.5	89.4	108.9	112.4	77.1	106.7
(Tung 2017)	77.6	91.4	89.9	88	107.3	110.1	75.9	107.5
PEPE-4V	77.6	85.3	82.8	93.9	94	101.1	83	102.7
PEPE-6V	67.5	72	66.8	69.2	72.1	77.1	65.1	68.4
PEPE-8V	**52.7**	**58.6**	**56.1**	67.3	**57.9**	**61.4**	**49.7**	**51.5**
	Sit	**SitD**	**Smoke**	**Wait**	**Walk**	**WalkD**	**WalkT**	**Avg**
(Luo 2018)	88.2	103.4	73.8	72.1	83.9	58.1	65.4	76
(Pavlakos 2017)	83.7	96.5	71.7	65.8	74.9	59.1	63.2	71.9
(Zhou 2017)	75.2	111.6	64.2	66.1	63.2	51.4	55.3	64.9
(Martinez 2017	81.9	**99.8**	69.1	63.9	50.9	67.1	54.8	67.5
(Wandt 2019)	100.5	125.8	88	84.8	72.6	78.8	79	89.9
(Wu 2016)	127.4	139	103.4	91.4	79.1	-	-	98.4
(Tung 2017)	124.2	137.8	102.2	90.3	78.6	-	-	97.2
PEPE-4V	100.6	125.9	88.1	84.9	72.7	78.9	79.1	90.1
PEPE-6V	83.8	107.5	71.8	65.9	75	59.2	63.3	72.9
PEPE-8V	**71.1**	99.9	**60.1**	**62**	**59.1**	**47.3**	**51.2**	**61**

Figure 4: We show reconstruction for variety of motions from the test set of Human3.6M. Six samples are shown. The pictures from left to right are input, ground truth and our pose estimation (PEPE-8V).



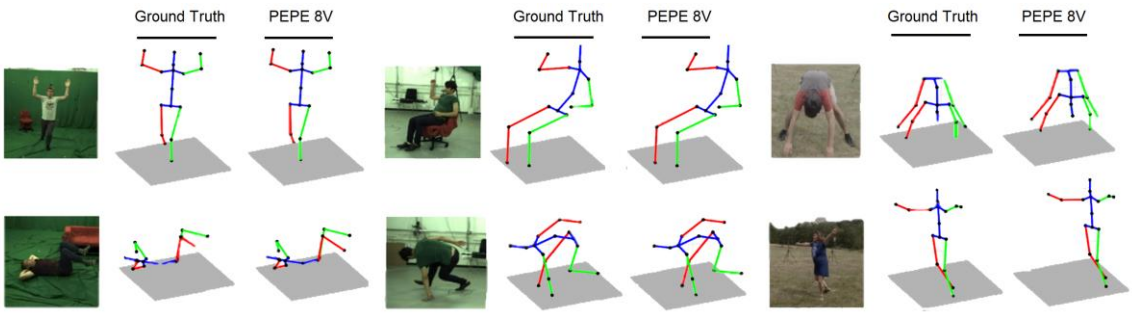
Although Human3.6M is outside our training set, we still outperform models that are trained on this data, and thereby highlights the generalization ability of our approach. For subjective evaluation, we compare the same motion sequence from the same viewing angle in figure 4. Challenging poses such as seated with crossing legs are included.

Although our method is trained using Panoptic studio dataset only, we outperform supervised and unsupervised approaches trained on Human3.6M dataset in many scenarios. Our results for MPII outperform most of the methods that are trained on this dataset without additional training. We report quantitative results in Table 2. Figure 5 shows prediction samples.

Table 2: *Comparison of our method against the state of the art on single person MPII test set. All evaluations use ground-truth bounding box crops around the subject. We report the PCK measure in 3D.*

Method	Stand/ Walk	Exercise	Sit on Chair	Crouch/ Reach	On the Floor	Sport	Misc.	Avg
(Mehta 2018)	85.7	75.4	**78.6**	72.3	**60.2**	81.8	73.4	75.3
(Mehta 2017, A)	86.6	75.3	74.8	73.7	52.2	82.1	77.5	74.6
(Mehta 2017, B)	87.7	77.4	74.7	72.9	51.3	83.3	**80.1**	75.3
(Zhou 2017)	85.4	71	60.7	71.4	37.8	70.9	74.4	67.4
PEPE-4V	72.4	58.2	60.4	71.3	41.5	59.6	59.5	60.4
PEPE-6V	82.0	71.7	74.9	68.6	56.5	78.1	69.7	71.7
PEPE-8V	**88.45**	**80.15**	77.65	**81.55**	60.05	**83.95**	79.3	**76.4**

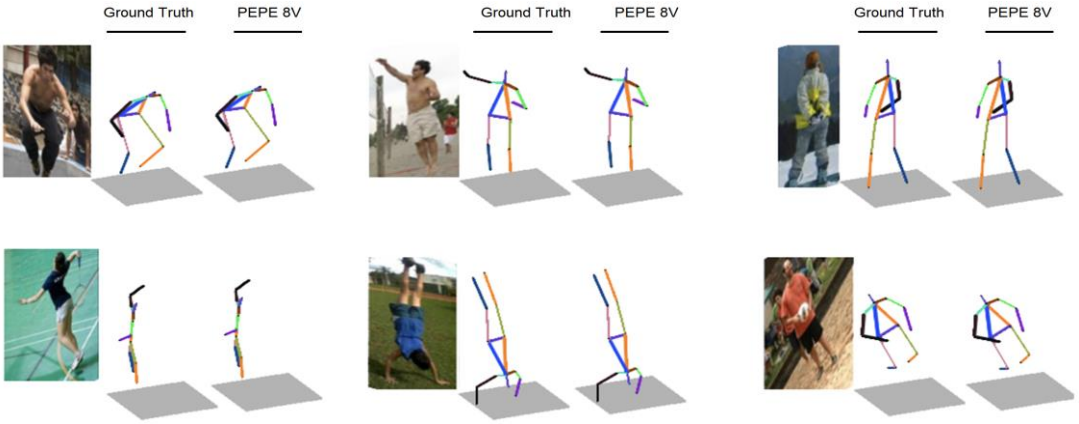
Figure 5: Variety reconstructions of motions from test set of MPII created using our method with 8 views.



Qualitative Evaluation

LSP is commonly used for qualitative evaluation due to its small content yet sparse characteristics. With 2000 images of humans during sport activities with some of these poses were never seen by our network. Despite this fact, our method reconstructed poses as shown in figure 3 and predicted plausible 3D poses for many images as shown in figure 6.

Figure 6: Variety reconstructions of motions from test set of LSP created using our method with 8 views.



The tests cover cases that are captured from uncommon camera angles. The third column in figure 6 illustrates cases in which the model failed to generalize. Table 3 shows LSP results.

Table 3: PCK-based comparison with state of the art on the LSP test set. We evaluate using 4,6 and 8 views.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average
(Rafi 2016)	95.8	86.2	79.3	75	86.6	83.8	79.8	83.8
(Belag. 2017)	95.2	89	81.5	77	83.7	87	82.8	85.2
(Lifshit 2016)	96.8	89	82.7	79.1	90.9	86	82.5	86.7
(Pishlin 2016)	97	91	83.8	78.1	91	86.7	82	87.1
(Insaf 2016)	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
(Wei 2016)	97.8	92.5	87	83.9	91.5	90.8	89.9	90.5
(Email 2019)	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
(Chu 2017)	98.1	93.7	89.3	86.9	93.4	94	92.5	92.6
(Yang 2017)	98.3	94.5	92.2	88.9	94.7	95	93.7	93.9
(Ning 2018)	98.2	94.4	91.8	89.3	94.7	95	93.5	93.9
(Chou 2018)	98.2	94.9	92.2	89.5	94.2	95	94.1	94
(Bulat et al., 2020)	98.7	95.7	93.1	90.3	95.8	95.6	94.8	94.8
PEPE-4V	96.16	89.96	82.46	77.96	84.66	87.96	83.76	86.16
PEPE-6V	96.7	92.3	87.9	85.5	92.0	92.6	91.1	91.2
PEPE-8V	**99.3**	**96.5**	**93.9**	**91.1**	**96.6**	**96.4**	**95.6**	**95.6**

5. Conclusion

In this paper, we propose a 3D body pose estimation method by evaluating the shared latent space posterior of the depth map and body pose parameters. We used two deep generative networks: a variational autoencoder to generate multiple camera views of body poses and a generative adversarial network to model the prior of body poses and depth mapping for the poses. This results in the ability to exploit the generalization properties of the GAN as well as the pose constraints implicitly learned by the VAE to improve discriminative pose estimation.

The proposed method learns from unlabeled data, which overcomes a significant problem in the field of body pose estimation, where annotated training data is scares. Our approach enhances the semi-supervised configurations of GAN to make more structured predictions. Our approach steadily results in better performance and generalization against three unseen datasets over previous semi-supervised and unsupervised state of art methods.

References

A. Bulat, J Kossaifi, G Tzimiropoulos, M Pantic. (2020) Toward fast and accurate human pose estimation via soft-gated skip connections. *In IEEE FG.*

A. Kanazawa, M. J. Black, D. W. Jacobs, J. Malik. (2018) End-to-end recovery of human shape and pose. *In CVPR.*

A. Popa, M. Zanfir, C. Sminchisescu. (2017) Deep Multi-task Architecture for Integrated 2D and 3D Human Sensing. *In CVPR.*

B. Tekin, A. Rozantsev, V. Lepetit, P. Fua. (2016) Direct prediction of 3d body poses from motion compensated sequences. *In CVPR.*

B. Tekin, P. Marquez-Neila, M. Salzmann, P. Fua. (2017) Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. *In ICCV.*

B. Wandt, H. Ackermann, and B. Rosenhahn. (2018) A kinematic chain space for monocular motion capture. *In ECCV.*

B. Wandt, Bodo Rosenhahn. (2019) Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. *In CVPR.*

B. Email, T. zimiropoulos. (2016) Human Pose Estimation via Convolutional Part Heatmap Regression. *In ECCV.*

C. Chou, J. Chien, H. Chen. (2018) Self Adversarial Training for Human Pose Estimation. *In APSIPA.*

C. Ionescu, D. Papava, V. Olaru, C. Sminchisescu. (2014) Human3.6m: Large scale datasets and predictive methods for 3dhuman sensing in natural environments. *In Transactions on Pattern Analysis and Machine Intelligence.*

C. Szegedy. (2015) Going deeper with convolutions. *In CVPR.*

C. Luo, X. Chu, A. L. Yuille. (2018) Orinet: A fully convolutional network for 3d human pose estimation. *In BMVC.*

C. Wan, T. Probst, L. Van Gool, A. Yao. (2017) Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation. *In CVPR.*

C. Ionescu, O. Vantzos, C. Sminchisescu. (2015) Matrix Backpropagation for Deep Networks with Structured Layers. *In ICCV.*

C. Chen, A. Tyagi, A. Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, James M.Rehg. (2019) Unsupervised 3d pose estimation with geometric self-supervision. *In CVPR.*

D. Mehta, H. Rhodin, D. Casas, P. Fua, O. Sotnychenko,W. Xu, C. (2017) Theobalt. Monocular 3d human pose estimation in the wild using improved CNN supervision. *In 3D Vision.*

D. Mehta, O. Sotnychenko, F. Mueller, W. Xu, S. Sridhar, G. Pons-Moll, C. Theobalt. (2018) Single-shot multi-person 3dpose estimation from monocular RGB. *In 3D Vision.*

D. Mehta, S. Sridhar, O. Sotnychenko, H. Rhodin, M. Shafiei, H.-P. Seidel, W. Xu, D. Casas, C. Theobalt. (2017) Vnect: Real-time 3d human pose estimation with a single RGB camera. *In Transactions on Graphics.*

D. Novotny, N. Ravi, B. Graham, N. Neverova, A. Vedaldi. (2019) C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure from Motion. *In ICCV.*

D. Pavllo, C. Feichtenhofer, D. Grangier, M. Auli. (2019) 3D Human Pose Estimation in Video with Temporal Convolutions and Semi-Supervised Training. *In CVPR.*

D. Tome, C. Russell, L. Agapito. (2018) Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. *In ICBEB.*

E. Simo-Serra, A. Ramisa, G. Aleny, C. Torras, F. MorenoNoguer. (2012) Single image 3d human pose estimation from noisy observations. *In CVPR.*

F. MorenoNoguer. (2017) 3d human pose estimation from a single image via distance matrix regression. *In CVPR.*

G. Moon, J. Y. Chang, K. M. Lee. (2019) Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation from a Single RGB Image. *In ICCV.*

G. Ning, Z. Zhang, Z. He. (2018) Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation. *In Transactions on Multimedia.*

G. Pavlakos, X. Zhou, K. G. Derpanis, K. Daniilidis. (2017) Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations. *In CVPR.*

G. Pavlakos, X. Zhou, K. G. Derpanis, K. Daniilidis. (2017) Coarse-to-fine volumetric prediction for single-image 3d human pose. *In CVPR.*

G. Rogez, Cordelia Schmid. (2016). MoCap-guided data augmentation for 3D pose estimation in the wild. *In NIPS.*

H. F. Tung, A. W. Harley, W. Seto, K. Fragkiadaki. (2017) Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. *In ICCV.*

H. Joo. (2015) Panoptic Studio: A Massively Multiview System for Social Motion Capture. *In ICCV.*

I. Habibie, W. Xu, D. Mehta, G. Pons-Moll t, C. Theobalt. (2019) In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. *In CVPR.*

I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu,D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio. (2014) Generative adversarial nets. *In NIPS.*

I. Insafutdinov, E. Eldar. (2016) DeeperCut: A Deeper, Stronger, and Faster Multi-person Pose Estimation Model. *In ECCV.*

J. Martinez, R. Hossain, J. Romero, J. J. Little. (2017) A simple yet effective baseline for 3d human pose estimation. *In ICCV.*

J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, W. T. Freeman. (2016) Single image 3d interpreter network. *In ECCV.*

K. Kingma, Ba, J. (2015). Adam: A Method for Stochastic Optimization. *In ICLR.*

L. Pishchulin. (2016) DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. *In CVPR.*

L. Lifshitz, Fetaya, Ullman. (2016) Human Pose Estimation Using Deep Consensus Voting. *In ECCV.*

M. Arjovsky, S. Chintala, L. Bottou. (2017) Wasserstein generative adversarial networks. *In ICML.*

M. Kocabas, S. Karagoz, E. Akbas. (2019) Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. *In CVPR.*

M. Rayat Imtiaz Hossain, J. J. Little. (2018) Exploiting temporal information for 3d human pose estimation. *In ECCV.*

P. Diederik, Kingma, Max Welling. (2014) Auto-Encoding Variational Bayes. *In ICLR.*

R. Rafi, Leibe, Gall, Kostrikov. (2016) An Efficient Convolutional Network for Human Pose Estimation. *In BMVC.*

S. Johnson, M. Everingham. (2010) Clustered pose and nonlinear appearance models for human pose estimation. *In BMVC.*

S. Park, J. Hwang, N. Kwak. (2016) 3d human pose estimation using convolutional neural networks with 2d pose information. *In ECCV Workshops.*

S. Wei, V. Ramakrishna, T. Kanade, Y.Sheikh. (2016) Convolutional Pose Machines. *In CVPR.*

V. Belagiannis, A. Zisserman. (2017) Recurrent Human Pose Estimation. *In IEEE FG.*

W. Yang, S. Li, W. Ouyang, H. Li, X. Wang. (2017) Learning Feature Pyramids for Human Pose Estimation. *In ICCV.*

X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, X. Wang. (2017) Multi-context Attention for Human Pose Estimation. *In CVPR.*

X. Zhou, Q. Huang. (2017) Towards 3d human pose estimation in the wild: A weakly supervised approach. *In ICCV.*

X. Sun, J. Shang, S. Liang, Y. Wei. (2017) Compositional Human Pose Regression. In ICCV.

X. Zhou, Q. Huang, X. Sun, X. Xue, Y. We. (2017) Weakly-Supervised Transfer for 3D Human Pose Estimation in the Wild. *In ICCV.*