Generative Multi-View Based 3D Human Pose Estimation

Motaz Sabri
Ridge-I
Tokyo, Japan
msabri@ridge-i.com

ABSTRACT

Large amounts of annotated data is essential for modern Human pose estimation. We propose using a semi supervised learning scheme to estimate the 3D poses from adversarial multi-views generated representations from a single RGB image. Our GAN generated views are the result of training that aims to create authentic and less degenerated outputs. Our method targets the shared latent space between the 3 dimensional and 2 dimensional poses and aims to simplify it by constraining the latent distribution. This resulted in a noticeable increase in the method generalization and exploitation of unlabeled depth maps. We utilized heatmaps to visualize the attention robustness under variety of poses. Our method competes with state of the art performances among semi supervised approaches and excels in some challenging poses as evaluated on Human3.6M, MPII-INF-3DHP and Leeds SportsPose challenging datasets. [1]

CCS CONCEPTS

• **Theory of computation** → **Semi-supervised learning**.

KEYWORDS

Panoptic reconstruction, View generation, 3D reconstruction, inpainting, VAE

ACM Reference Format:
Motaz Sabri. 2021. Generative Multi-View Based 3D Human Pose Estimation. In *SIET '21:International Conference on Sustainable Information Engineering and Technology, September 13–14, 2021,Malang, Indonesia.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3479645.3479708

1 INTRODUCTION

3D human pose estimation refers to the detection and localization of the human body joints in videos and images. It offers key information to analyzing human behavior such as human-robot interaction and action recognition. There are numerous approaches to handle generating 3D human poses from monocular images [10, 17, 34–37, 39, 42, 43, 53]. 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. However, These methods are still limited to the poses similarity between training and testing samples and therefore tend to have degraded quality. The training data distribution has dominant effect on the model behavior which limits its generalization abilities toward unseen views.

Semi-supervised learning provides an alternative method for learning robust geometry representations without extensive precise 3D annotation. Many approaches [4, 16, 28, 38, 45, 63] 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.

On one hand, using the semi supervised methods to exploit information from unlabeled data is highly non-trivial. On the other hand, overcoming the generalization challenge of supervised methods requires a great deal of annotation, which is tedious and error-prone. In this work, we train an adversarial multi-view images generator from original RGB image to provide semi supervision and estimate 3D human pose. We utilize a variational autoencoder (VAE) [9] generative model to create the views and generative adversarial network (GAN) [14] to capture the latent spaces of human body poses and corresponding images for estimating 3D pose. We also propose a mapping between the latent body pose space and latent joint-mapping space. A key point that is sampled in any of those latent spaces is defined by the VAE as a 3D pose or by GAN as a joint mapping. As a result, the model learns geometrical representations that enables decent 3D poses estimation. During inference, our generator creates different views of the given target which adopts the model parameters to the new distribution. This adaptation enables the model to generalize well in estimating 3D poses for unseen samples.

Motivated by Self-Attention GAN's [15] (SAGAN), we incorporate them to enhance the 3D pose estimation. It offers global structural constrains for the body joints through learning the general and instantaneous dependencies between those joints. Our results steadily showed the effectiveness of our method on Human3.6M [19], MPII-INF-3DHP (MPII) [36], and Leeds Sports Pose (LSP) [21] datasets. Besides drastically reducing the need for annotated data, our approach enhances the performance under scaling and viewpoint variation. Figure 1 shows an overview of the proposed method.

Our paper is organized as follows. Related works is reviewed in section 2. Proposed method is presented in Section 3. Experimental results are discussed in Sections 4, followed by the conclusion in Section 5.

2 RELATED WORKS

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

[1] This is an extended and revised version of a preliminary conference report that was presented in KICONF 2020 [54]

SIET '21, September 13–14, 2021, Malang, Indonesia

ACM ISBN 978-1-4503-8407-0/21/09…$15.00
https://doi.org/10.1145/3479645.3479708

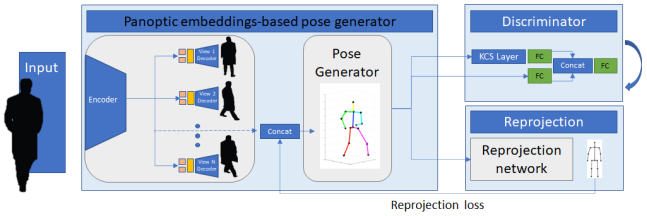

Figure 1: The proposed architecture consists of three parts: Panoptic embeddings based pose generator: 3D pose estimator that uses features extracted from a multi view VAE, Discriminator with wasserstein loss that scores the realness of a given image instead of the probability of the image realness. Lastly the reprojection network guided by [63].

for training [35, 37, 43, 44, 48, 53, 57, 59, 60, 74]. 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 [22, 36, 44] and data augmentation [20, 27, 52]. Transfomer based models have benefited from such increase but without sufficent generalization [33, 72]. However, the limited diversity of appearance and motion that current tools provide, along with their imperfect verisimilitude, limits both the generality and the accuracy of networks trained using only those images.

Due to the previous factors, semi supervision is considered a promising alternative in which network is trained without 2D to 3D supervision [4, 16, 28, 38, 45, 63]. Researchers utilized semi supervision to let their models acquire knowledge through different perspectives. Kocabas et al. [28] propose Epipolar geometry to self-supervise a 3D pose estimator. Bastian et al. [63] used reporejoction loss to train a semi supervised model. They also use adversarial losses to identify between infeasible poses. Kanazawa et al. [24] create a parameterized 3D mesh with respect to joint angles in 3D space and the linear shape space. Pavllo et al. [45] utilize temporal information with dilated convolutions over 2D key point trajectories to estimate 3D poses in videos which is found helpful when labeled data is scarce. The robustness of 3D pose estimation has increased through the above proposals however there is still space to improve generalization against viewpoints and scale changes. Several geometry-driven self-supervised methods [4, 11, 28, 29, 41, 46, 50, 51, 60, 61] are proposed to train models with more unlabeled training data. However, they were not evaluated on active manner. Tome et al. [60] use motion capture data and multi-staged 2D pose estimation model to generate 2D and 3D pose predictions. Novotny et al. [41] Estimate 3D pose using viewpoint and shape parameters and used canonicalization loss to introduce inductive bias. Wan et al. [61] learn to map two latent spaces by using two generative models. Their models generation abilities allowed the pose constraints to be learned improving the pose estimation.

In this paper, we propose a semi-supervised method that is trained with view generator from a single RGB image. Our work is aligned with [28, 44] in using the generated data for semi supervised training. The approach of [44], however, calibrates a set of camera views parameters that are hard to obtain in open environments. The [28] approach uses Epipolar geometry to estimate the corresponding 3D pose from 2D poses. This results in back propagating the 3D reconstruction errors to the trained models and consequentially builds against outdoor environments where 2D pose estimation is not robust. In contrast to [44] and [28], our learning which uses two generative models is robust to challenging poses in the captured data and unbound to view constrains which concurrently optimize for 2D and 3D poses. Our learning also improves predictions in 2D space using multi-view unlabeled data. We evaluate our approach on three challenging datasets unseen during training and find out that it competes with existing methods trained on these datasets for semi supervised learning and excels in the 3D pose estimation task under some challenging poses.

3 PROPOSED METHOD

We propose a framework that maps from the input 2D 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. GAN [14] 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). The errors in the 3D poses from input 2D observations is minimized by adopting the Wasserstein loss for the GAN [1, 63]. Figure 1 shows the network main components: 1) the generator network, 2) the discriminator used in the adversarial training and finally 3) the reprojection part that maps the 2D poses distribution to the 3D poses distribution. The three components are trained alternatively.

3.1 Generator

The generator models a prior distribution on body pose configurations using inception [58] VAE. Its structure allows learning the mapping from high dimensional body poses to a low-dimensional representation while keeping the reconstruction accuracy high through 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

$$p(g) = \int_h p(g \mid h)p(h)dh \tag{1}$$

Fitting $p(g)$ directly usually involves expensive inference. Therefore, we approximate $p(g)$ using VAE guided by a GAN. We provide a brief mathematical representation for the usage of VAE and GAN and their usage to model the prior of depth mapping and body 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. $\overline{y}$ refers to the synthesized depth map from the GAN generator. h_{x_n} and h_{y_n} are the n^{th} latent pose of view and depth map respectively.

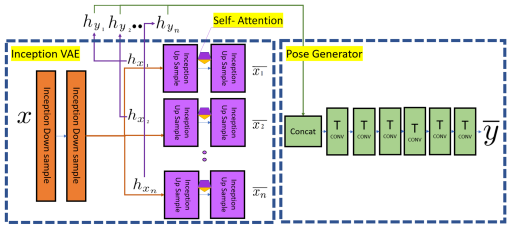

Figure 2: The panoptic embeddings based pose generator (referenced in Figure 1) is shown. The VAE component and the generator create authentic $\overline{y}$ to fool the discriminator.

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

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

$$Dec(h_{x_n}) = p(x \mid h_{x_n}) \tag{3}$$

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 D_{KL} between the encoded distribution $q(h_{x_n} \mid x)$ and the prior $p(h_{x_n})$. The VAE loss is given by reconstruction errors summation of N decoders and latent prior as follow:

$$\mathcal{L}_{vae} = \sum_{n=1}^{N} \mathcal{L}_n^{pose} + \mathcal{L}_p \tag{4}$$

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

$$\mathcal{L}_n^{pose} = -\mathbb{E}_{q(h_{X_n}|x)} \left[\log p(x|h_{x_n})\right] \tag{5}$$

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

$$\mathcal{L}_p = \sum_{n=1}^{N} D_{KL} \left(q(h_{x_n} \mid x) \,||\, p(h_{x_n})\right) \tag{6}$$

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. By optimizing latent space parameter of the body pose, we create a reference space to learn mapping to the 3D map latent space that is H_y through $MAP(H_x)$.

Having the corresponding pairs x and y, we can train using the 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 minimize the 3D reconstruction error by the intermediate loss $\mathcal{L}_r$ given a latent input H_x which is mapped to the GAN latent space as follows:

$$\mathcal{L}_r = \max\left(\parallel y_n - Gen(MAP(H_x)) \parallel^2, \xi \right) \tag{7}$$

Where ξ is the clipping threshold and $MAP(\cdot)$ 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 [55]. Since the depth map is normalized to $[1, -1]$, we set $\xi = 1$. Since our generative model is able to learn low-dimensional representations, we are able to generate realistic samples with a very small set of labeled (x, y) pairs. With such mapping, arbitrary points in the latent pose space can be mapped into the 2D body pose space or corresponding 3D map space. The latent spaces of this mapping is considered as a common shared latent pose.

The composite function $Gen(MAP(\cdot))$ generates the depth latent space. Its input is the mapped latent space marked by the green arrow in Figure 2. The normal distribution H_x is mapped to H_y through the $MAP(\cdot)$ that is implicitly learned. This infers that any standard normal distribution random noise can be mapped to body pose or a corresponding depth map. The $Gen(MAP(\cdot))$ function is implemented as six transposed convolutional layers with dilation factor of two in order to build $\overline{y}$. A shared latent space between the depth map and the pose can be simplfied by introducing the constraint on the latent distribution. The generated samples $\overline{y}$ and real data samples y are fed to the discriminator to distinguish between them.

3.2 Attention Localization

Following the self-attention mechanism mentioned in [15] which indicates that, for an arbitrary input, the overall self-attention map is in size of (H × W) × (H × W). Therefore, there is a corresponding sub-attention map whose size is (H × W) for every specific point in the image [11]. We incorporated the self-attention layer in the last two layers of decoders as shown in Figure 2. This allows our model to generate representations using salient details from all feature positions. The discriminator can observe that distant features in a single image are consistent with each other. Following [15] to enhance feature learning, The GAN generator uses spectral normalization. Figure 3 shows attention layer visualization. It is observed that the generator enhances neighborhoods that correspond to joints locations in all views.

3.3 Discriminator

We chose to design the discriminator to be similar to the pose regression network. We train it using the Wasserstein loss function [1]. We incorporate kinematic chain space (KCS) [62] to handle joint angle dynamics, kinematic chains, and symmetry. It operates based on the constraint that bone lengths are constant. This results in trivial optimization of a nuclear norm 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. The KCS branch output is concatenated with another fully connected layer from the pose generator. Consecutively, the GAN loss with respect to generator and discriminator is:

$$\mathcal{L}_{GAN} = \log(Disc(y)) + \log(1 - Disc(Gen(H_y))) \tag{8}$$

where $Disc(y)$ is the discriminator output and $Gen(H_y)$ is the generated pose. The generator learns to reduce the loss and generates more authentic samples while the discriminator is trained to confuse the generator by maximizing the loss.

3.4 Reprojection

The reprojection layer takes as input the synthetic poses $\overline{y}$ created by the generator and remaps it into 2D coordinate space. This proposal follows the method of [63]. It is given as:

$$w' = y \tag{9}$$

where w' denotes the 2D reprojection. The reprojection loss function $\mathcal{L}_j(\cdot)$ can be defined as:

$$\mathcal{L}_j(\overline{y}) = \| \ w - \overline{y} \ \|_F \tag{10}$$

Where $\| \ . \ \|_F$ represents the Frobenius norm and w is the observed 2D pose matrix which has the same structure as w'. The reprojection layer does not have any 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. According to the discriminator, the pose generator will hallucinate the missing joints.

4 EXPERIMENTS

4.1 Training

We trained our model using a fine-grained subset of the Panoptic studio [22] dataset. It contains 480 synchronized video streams for numerous subjects involved in social activities with their anatomical landmarks labels. 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. Our selected streams are for scenes that have no 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. Our selected samples have a variety of person sizes which gave the 2D joint generator an arbitrary scale. To minimize this effect, The vector of the generated 2D pose is divided by its own 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 three components of our model 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 [26] algorithm to optimize our network. We regularized latent variable after VAE encoder by adding random Gaussian noise with 0.035 standard deviation. The training took 5000 epochs with learning rate set to 0.001. It took about 15 hours for training with around 35,000 samples on 4X NVIDIA Tesla V100 of 32GB.

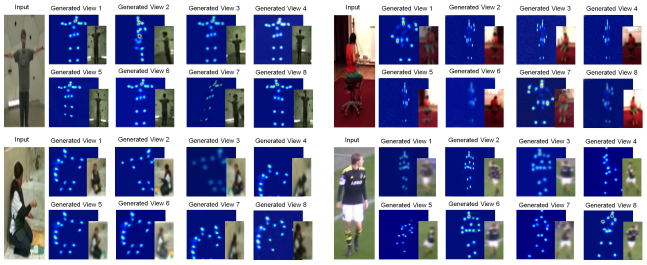

Figure 3: The PEPE is able to reconstruct 8 views that correspond to the input image. We include the self-attention map referenced in 2. The subject joints in the corresponding view is localized under variation of their activation. They hold the desired structural information. The top left image is from Panoptic studio (testing data). The top right image is from Human3.6M. The input in this case is not the front view. However, the model managed to make successful reconstruction. Image at bottom left and bottom right are from the MPII and LSP datasets respectively.

4.2 Evaluation

After training our model with Panoptic studio dataset [22], we evaluate its performance on the following three datasets: Human3.6M [19], MPII [36], and LSP [21]. Human3.6M dataset contains images of people aligned with respect to time into 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, we evaluate on LSP that contains unusual poses and camera angles.

As multi-view generation is an intermediate step and its effect can not be seen directly through results. Figure 3 shows the PEPE reconstruction of 8 views from a single RGB image. The inputs are unseen data with a variety of postures and and viewing points. The generated views hold useful structural information that can be used for pose estimation once projected on the depth map space. In Figure 3, for every input image, we show the decoded view and the generated attention for every view. The positional attention module captures structural joint information. For comprehensible visualization, we show some attended channels that highlights clear joint information. The activation around a specific join is observed after enhancing the channel attention module. These visualizations show ability to capture dependencies for improving feature representation in 2D pose estimation.

4.2.1 Quantitative Evaluation. Quantitative analysis in the literature uses the mean per joint positioning error (MPJPE) and Percentage of Correct Key points (PCK). The MPJPE aims to compute the error per joint position as the Euclidean distance between prediction and the ground truth of a given 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 followed same conventions in our analysis. In

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

Method	Direct	Disc	Eat	Greet	Phone	Photo	Pose	Perch	Sit	SitD	Smoke	Wait	Walk	WalkD	WalkT	Avg
Martinez et al.[35]	51.8	56.2	58.1	59.0	69.5	78.4	55.2	58.1	74.0	94.6	62.3	59.1	65.1	49.5	52.4	62.9
Iskakove et.al [25]	41.9	49.2	46.9	47.6	50.7	57.9	41.2	50.9	57.3	74.9	48.6	44.3	41.3	52.8	42.7	49.9
Pavllo et al. [45]	45.2	46.7	43.3	45.6	48.1	55.1	44.6	44.3	57.3	65.8	47.1	44.0	49.0	32.8	33.9	46.8
Cheng et al.[70]	38.3	41.3	46.1	40.1	41.6	51.9	41.8	40.9	51.5	58.4	42.2	44.6	41.7	33.7	30.1	42.9
Cheng et al.[6]	**36.2**	38.1	42.7	**35.9**	38.2	**45.7**	**36.8**	42.0	45.9	**51.3**	**41.8**	**41.5**	43.8	33.1	**28.6**	**40.1**
Dinehs et al.[75]	38.4	46.2	44.3	43.2	44.8	48.3	52.9	36.7	**45.3**	54.5	63.4	44.4	41.9	46.2	39.9	44.6
Mohamaddi et al.[23]	39.4	46.9	41.0	42.7	53.6	54.8	41.4	50.0	59.9	78.8	49.8	46.2	51.1	40.5	41.0	49.1
Fang et al.[13]	50.1	54.3	57.0	57.1	66.6	73.3	53.4	55.7	72.8	88.6	60.3	57.7	62.7	47.5	50.6	60.4
Zhao et al.[73]	47.3	60.7	51.4	60.5	61.1	49.9	47.3	68.1	86.2	55.0	67.8	61.0	42.1	60.6	45.3	57.6
Cai et al.[67]	44.6	47.4	45.6	48.8	50.8	59.0	47.2	43.9	57.9	61.9	49.7	46.6	51.3	37.1	39.4	48.8
Xu et al.[66]	37.4	43.5	42.7	42.7	46.6	59.7	41.3	45.1	52.7	60.2	45.8	43.1	47.7	33.7	37.1	45.6
liu et al.[32]	41.8	44.8	41.1	44.9	47.4	54.1	43.4	42.2	56.2	63.6	45.3	43.5	45.3	31.3	32.2	45.1
Zeng et al.[72]	46.6	47.1	43.9	41.6	45.8	49.6	46.5	40.0	53.4	61.1	46.1	42.6	43.1	31.5	32.6	44.8
Wang et al.[32]	40.2	42.5	42.6	41.1	46.7	56.7	41.4	42.3	56.2	60.4	46.3	42.2	46.2	31.7	31.0	44.5
PEPE-4V	42.2	49.4	47.7	58.3	58.1	65.7	47.8	67.5	64.6	90.7	53	49.5	37.5	43.3	43.5	54.6
PEPE-6V	40.6	48.1	46.7	57.1	56.4	64.5	46	66.1	62.7	88.8	51.7	48	36.5	41.7	42.4	53.1
PEPE-8V	38.8	**37.9**	**39.8**	45.9	**38.1**	51.6	37.1	**36.6**	62	87.2	43.4	42.1	**34.4**	**30.3**	29.1	43.6

our evaluation we use Protocol-1 where no pose alignment happens. Table 1 shows our results on the Human3.6M dataset. We observe that having more generated views significantly improves the pose estimation. Our method competes with supervised and unsupervised methods in many cases, namely [6] and [75] and excels in some challenging poses such as perching and sitting down. 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. To keep the evaluation subjective, we analyze the same movement from the same viewing point in Figure 4. We included standard poses and challenging poses such as crouching and squatting.

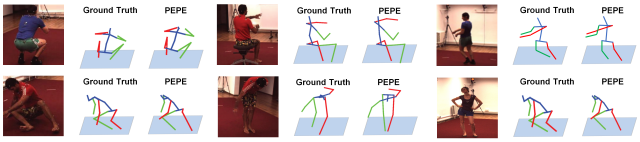

Figure 4: We show reconstruction for variety of motions from the test set of Human3.6M. Six samples are shown. The images starting from the left are input, ground truth and our pose estimation (PEPE-8*V*).

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 compete with the state of the art methods that are trained on this dataset without additional training namely the work of Bulat et al.[3]. Figure 5 shows some standard and challenging prediction samples. We report quantitative results in Table 2.

4.2.2 Qualitative Evaluation. LSP is commonly used for qualitative evaluation due to its small content yet sparse characteristics.

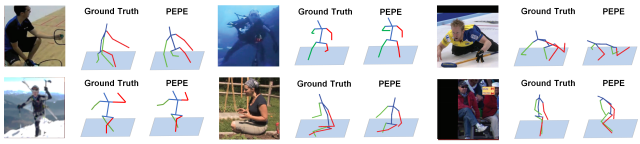

Figure 5: We show reconstruction for variety of motions from the test set of MPII created by considering 8 views (PEPE-8*V*).

This dataset contains 2000 images of people showing variety of poses while performing sports activities. Our network never encountered many of the evaluated poses during training. 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. 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 the results of LSP dataset outperforming all the state of the art methods for this dataset. This underlines robustness of the learning to distinguish between feasible and infeasible poses and 2D projections.

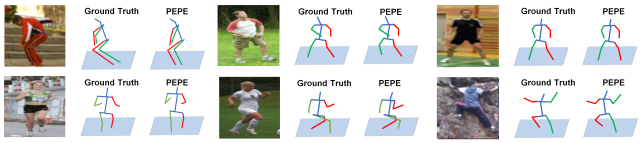

Figure 6: Reconstructions is shown for a variety of motions from the test set of LSP created by considering 8 views (PEPE-8*V*).

Table 2: Comparison on MPII test set with the state of the art. Bounding box around people are used for evaluation. We report the PCK measure in 3D. The higher the score the more accurate is the estimation.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average
Chen et al.[5]	98.1	96.5	92.9	88.5	90.2	89.6	86.0	91.9
Ke et al.[31]	98.5	96.8	92.7	88.4	90.6	89.4	86.3.5	92.1
Tang et al.[65]	98.4	96.9	92.6	88.4	91.8	89.4	86.2	92.3
Sun et al. [56]	98.6	96.9	92.8	89.0	91.5	89.0	85.7	92.3
Bulat et al.[3]	**98.8**	97.5	94.4	**91.2**	**93.2**	92.2	**89.3**	**94.1**
PEPE-4V	88.45	80.15	77.65	81.55	60.05	83.95	79.35	78.7
PEPE-6V	97.2	96.5	92.1	87.3	90.9	87.9	84.6	90.9
PEPE-8V	98.7	**98.1**	**95.4**	91.1	92.1	**92.5**	88.9	93.8

Table 3: PCK-based comparison with state of the art on the LSP test set. We report PCK measure in 3D using 4, 6 and 8 views. Higher PCK indicates more accurate estimation.

Method	Head	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Average
Yang et al.[69]	90.6	78.1	73.8	68.8	74.8	69.9	58.9	73.6
Rafi et al.[49]	95.8	86.2	79.3	75	86.6	83.8	79.8	83.8
Yu et al.[71]	87.2	88.2	82.4	76.3	91.4	85.8	78.7	84.3
Belagiannis et al.[2]	95.2	89	81.5	77	83.7	87	82.8	85.2
Lifshitz et al.[30]	96.8	89	82.7	79.1	90.9	86	82.5	86.7
Pishchulin et al.[47]	97	91	83.8	78.1	91	86.7	82	87.1
Insafutdinov et al.[18]	97.4	92.7	87.5	84.4	91.5	89.9	87.2	90.1
Wei et al.[64]	97.8	92.5	87	83.9	91.5	90.8	89.9	90.5
Email et al.[12]	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al.[8]	98.1	93.7	89.3	86.9	93.4	94	92.5	92.6
Yang et al.[68]	98.3	94.5	92.2	88.9	94.7	95	93.7	93.9
Ning et al.[40]	98.2	94.4	91.8	89.3	94.7	95	93.5	93.9
Chou et al.[7]	98.2	94.9	92.2	89.5	94.2	95	94.1	94
Bulat et al.[3]	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. As a result, the learned pose constrains have improved the discriminative pose estimation by VAE and the GAN generalization . 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 results show competing performance against state of the art methods and surpass in some challenging poses. It also shows robust generalization against three unseen datasets over previous semi-supervised and unsupervised state of art methods.

REFERENCES

[1] M. Arjovsky, S. Chintala, and L. Bottou. 2017. Wasserstein generative adversarial networks. In *In International Conference on Machine Learning.*

[2] V. Belagiannis and A. Zisserman. 2017. Recurrent Human Pose Estimation. In *In International Conference on Automatic Faceand Gesture Recognition.*

[3] A. Bulat, J Kossaifi, G Tzimiropoulos, and M Pantic. 2020. Toward fast and accurate human pose estimation via soft-gated skip connections. In *In International Conference on Automatic Face and Gesture Recognition.*

[4] C. Chen, A. Tyagi, A. Agrawal, Dylan Drover, Rohith MV, Stefan Stojanov, and James M.Rehg. 2019. Unsupervised 3d pose estimation with geometric self-supervision. In *In Conference on Computer Vision and Pattern Recognition.*

[5] Y. Chen, C. Shen, X. Wei, L. Liu, and J. Yang. 2017. Adversarial PoseNet: A Structure-Aware Convolutional Network for Human Pose Estimation. In *IEEE International Conference on Computer Vision.*

[6] Y. Cheng, B. Yang, B. Wang, and R. T. Tan. 2020. 3D Human Pose Estimation Using Spatio-Temporal Networks with Explicit Occlusion Training. In *In Proceedings of the AAAI Conference on Artificial Intelligence.*

[7] C. Chou, J. Chien, and H. Chen. 2018. Self Adversarial Training for Human Pose Estimation. In *In Asia-Pacific Signal and Information Processing Association Annual Summit and Conference.*

[8] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. 2017. Multi-context Attention for Human Pose Estimation. In *In Conference on Computer Vision and Pattern Recognition.*

[9] P. Diederik, Kingma, and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations.*

[10] Y. Du, Y. Wong, Y. Liu, F. Han, Y. Gui, Z. Wang, M. Kankanhalli, and W. Geng. 2016. Marker-less 3D human motion capture with monocular image sequence and height-maps. In *In European Conference on Computer Vision.*

[11] D. Dylan, R. MV, C. Chen, A. Agrawal, A. Tyagi, and P. Cong. 2018. Can 3D Pose be Learned from 2D Projections Alone?. In *Proceedings of the European Conference on Computer Vision Workshops.*

[12] B. Email and T. zimiropoulos. 2016. Human Pose Estimation via Convolutional Part Heatmap Regression. In *In European Conference on Computer Vision.*

[13] H. Fang, Y. Xu, W. Wang, X. Liu, and S. Zhu. 2018. Learning pose grammar to encode human body configuration for 3d pose estimation. In *In AAAI Conference on Artificial Intelligence.*

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, D. Warde-Farley B. Xu, S. Ozair, A. Courville, and Y. Bengio. 2014. Generative adversarial nets. In *In International Conference on Neural Information Processing Systems.*

[15] D. Metaxas H. Zhang, I. Goodfellow and A. Odena. 2019. Self-attention generative adversarial networks. In *In International Conference on Machine Learning.*

[16] I. Habibie, W. Xu, D. Mehta, G. Pons-Moll t, and C. Theobalt. 2019. In the Wild Human Pose Estimation Using Explicit 2D Features and Intermediate 3D Representations. In *In Conference on Computer Vision and Pattern Recognition.*

[17] M. Rayat Imtiaz Hossain and J. J. Little. 2018. Exploiting temporal information for 3d human pose estimation. In *In European Conference on Computer Vision.*

[18] I. Insafutdinov and E. Eldar. 2016. DeeperCut: A Deeper and Stronger and and Faster Multi-person Pose Estimation Model. In *In European Conference on Computer Vision.*

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

[20] C. Ionescu, O. Vantzos, and C. Sminchisescu. 2015. Matrix Backpropagation for Deep Networks with Structured Layers. In *In International Conference on Computer Vision.*

[21] S. Johnson and M. Everingham. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *In British Machine Vision Conference.*

[22] H. Joo. 2015. Panoptic Studio: A Massively Multiview System for Social Motion Capture. In *In International Conference on Computer Vision.*

[23] A. Kadkhodamohammadi and N. Padoy. 2020. A generalizable approach for multi-view 3D human pose regression. In *CoRR, abs/1804.10462, 2018.*

[24] A. Kanazawa, M. J. Black, D. W. Jacobs, and J. Malik. 2018. End-to-end recovery of human shape and pose. In *In Conference on Computer Vision and Pattern Recognition.*

[25] I. Karim, B. Egor L. Victor, and M. Yury. 2019. Learnable triangulation of human pose. In *In International Conference on Computer Vision.*

[26] K. Kingma, Ba, and J. 2015. Adam: A Method for Stochastic Optimization. In *In International Conference on Learning Representations.*

[27] K. Knippenberg, J. Verbrugghe, I. Lamers, S. Palmaers, A. Timmermans, and A. Spooren. 2017. Markerless motion capture systems as training device in neurological rehabilitation: A systematic review of their use and application and target population and efficacy. In *In journal of NeuroEngineering and Rehabilitation.*

[28] M. Kocabas, S. Karagoz, and E. Akbas. 2019. Self-Supervised Learning of 3D Human Pose Using Multi-View Geometry. In *In Conference on Computer Vision and Pattern Recognition.*

[29] Li K. Jiang S. Zhang Z. Huang C. Xu R. Y. D. Li, Y. 2020. Geometry-Driven Self-Supervised Method for 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence.*

[30] L. Lifshitz, Fetaya, and Ullman. 2016. Human Pose Estimation Using Deep Consensus Voting. In *In European Conference on Computer Vision.*

[31] K. Lipeng, C. Ming-Ching, Q. Honggang, and L. Siwei. 2018. Multi-Scale Structure-Aware Network for Human Pose Estimation. In *European Conference on Computer Vision.*

[32] R. Liu, J. Shen, H. Wang, C. Chen, S. Cheung, and V. Asari. 2020. Attention mechanism exploits temporal contexts: Real-time 3d human pose reconstruction. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[33] A. Llopart. 2020. LiftFormer: 3D Human Pose Estimation using attention models. *arXiv preprint arXiv:2103.10455.*

[34] C. Luo, X. Chu, and A. L. Yuille. 2018. Orinet: A fully convolutional network for 3d human pose estimation. In *In British Machine Vision Conference.*

[35] J. Martinez, R. Hossain, J. Romero, and J. J. Little. 2017. A simple yet effective baseline for 3d human pose estimation. In *In International Conference on Computer Vision.*

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

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

[38] G. Moon, J. Y. Chang, and K. M. Lee. 2019. Camera Distance-Aware Top-Down Approach for 3D Multi-Person Pose Estimation From a Single RGB Image. In *In International Conference on Computer Vision.*

[39] F. MorenoNoguer. 2017. 3d human pose estimation from a single image via distance matrix regression. In *In the Conference on Computer Vision and Pattern Recognition.*

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

[41] D. Novotny, N. Ravi, B. Graham, N. Neverova, and A. Vedaldi. 2019. C3DPO: Canonical 3D Pose Networks for Non-Rigid Structure From Motion. In *In International Conference on Computer Vision.*

[42] S. Park, J. Hwang, and N. Kwak. 2016. 3d human pose estimation using convolutional neural networks with 2d pose information. In *In European Conference on Computer Vision Workshops.*

[43] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. 2017. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *In Conference on Computer Vision and Pattern Recognition.*

[44] G. Pavlakos, X. Zhou, K. G. Derpanis, and K. Daniilidis. 2017. Harvesting Multiple Views for Marker-Less 3D Human Pose Annotations. In *In Conference on Computer Vision and Pattern Recognition.*

[45] D. Pavllo, C. Feichtenhofer, D. Grangier, and M. Auli. 2019. 3D Human Pose Estimation in Video With Temporal Convolutions and Semi-Supervised Training. In *In Conference on Computer Vision and Pattern Recognition.*

[46] A. Pirinen, E. Gartner, and C. Sminchisescu. 2019. Domes to Drones: Self-Supervised Active Triangulation for 3D Human Pose Reconstruction. In *Advances in Neural Information Processing Systems.*

[47] L. Pishchulin. 2016. DeepCut: Joint Subset Partition and Labeling for Multi Person Pose Estimation. In *In Conference on Computer Vision and Pattern Recognition.*

[48] A. Popa, M. Zanfir, and C. Sminchisescu. 2017. Deep Multi-task Architecture for Integrated 2D and 3D Human Sensing. In *In Conference on Computer Vision and Pattern Recognition.*

[49] R. Rafi, Leibe, Gall, and Kostrikov. 2016. An Efficient Convolutional Network for Human Pose Estimation. In *In British Machine Vision Conference.*

[50] H. Rhodin, V. Constantin, I. Katircioglu, M. Salzmann, and P. Fua. 2019. Neural Scene Decomposition for Multi-Person Motion Capture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*

[51] H. Rhodin, M. Salzmann, and P. Fua. 2018. Unsupervised Geometry-Aware Representation for 3D Human Pose Estimation. In *Proceedings of the European Conference on Computer Vision.*

[52] G. Rogez and Cordelia Schmid. 2016. MoCap-guided data augmentation for 3D pose estimation in the wild. In *In International Conference on Neural Information Processing Systems.*

[53] G. Rogez, P. Weinzaepfel, and C. Schmid. 2017. LCR-Net: Localization-Classification-Regression for Human Pose. In *In Conference on Computer Vision and Pattern Recognition.*

[54] M. Sabri. 2020. Multi-view Generative Networks for 3D Pose Estimation. In *International Conference on Knowledge and Innovation in Engineering, Science and Technology.*

[55] E. Simo-Serra, A. Ramisa, G. Aleny, C. Torras, and F. MorenoNoguer. 2012. Single image 3d human pose estimation from noisy observations. In *In Conference on Computer Vision and Pattern Recognition.*

[56] K. Sun, B. Xiao, D. Liu, and J. Wang. 2018. Deeply Learned Compositional Models for Human Pose Estimation. In *European Conference on Computer Vision.*

[57] X. Sun, J. Shang, S. Liang, and Y. Wei. 2017. Compositional Human Pose Regression. In *In International Conference on Computer Vision.*

[58] C. Szegedy. 2015. Going deeper with convolutions. In *In Conference on Computer Vision and Pattern Recognition.*

[59] B. Tekin, P. Marquez-Neila, M. Salzmann, and P. Fua. 2017. Learning to Fuse 2D and 3D Image Cues for Monocular Body Pose Estimation. In *In International Conference on Computer Vision.*

[60] D. Tome, C. Russell, and L. Agapito. 2018. Lifting from the Deep: Convolutional 3D Pose Estimation from a Single Image. In *In International Conference on Biomedical Engineering and Bioinformatics.*

[61] C. Wan, T. Probst, L. Van Gool, and A. Yao. 2017. Crossing Nets: Combining GANs and VAEs with a Shared Latent Space for Hand Pose Estimation. In *In Conference on Computer Vision and Pattern Recognition.*

[62] B. Wandt, H. Ackermann, , and B. Rosenhahn. 2018. A kinematic chain space for monocular motion capture. In *In European Conference on Computer Vision.*

[63] B. Wandt and Bodo Rosenhahn. 2019. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *In Conference on Computer Vision and Pattern Recognition.*

[64] S. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh. 2016. Convolutional Pose Machines. In *In Conference on Computer Vision and Pattern Recognition.*

[65] T. Wei, Y. Pei, and W. Ying. 2018. Deeply Learned Compositional Models for Human Pose Estimation. In *European Conference on Computer Vision.*

[66] J. Wu, T. Xue, J. J. Lim, Y. Tian, J. B. Tenenbaum, A. Torralba, and W. T. Freeman. 2020. Deep kinematics analysis for monocular 3d human pose estimation. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[67] J. Liu J. Cai T. Cham Y. Cai, L. Ge. 2019. Exploiting spatial-temporal relationships for 3d pose estimation via graph convolutional networks. In *In Proceedings of the IEEE International Conference on Computer Vision.*

[68] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang. 2017. Learning Feature Pyramids for Human Pose Estimation. In *In International Conference on Computer Vision.*

[69] W. Yang, W. Ouyang, H. Li, and X. Wang. 2016. End-to-End Learning of Deformable Mixture of Parts and Deep Convolutional Neural Networks for Human Pose Estimation. In *In Conference on Computer Vision and Pattern Recognition.*

[70] C. Yu, Y. Bo, W. Bo, W. Yan, and T. Robby. 2019. Occlusion-aware networks for 3d human pose estimation in video. In *In International Conference on Computer Vision.*

[71] X. Yu, F. Zhou, and M. Chandraker. 2016. Deep deformation network for object landmark localization. In *In European Conference on Computer Vision.*

[72] A. Zeng, X. Sun, F. Huangand M. Liu, Q. Xu, and S. Lin. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *In European Conference on Computer Vision.*

[73] L. Zhao, X. Peng, Y. Tian, M. Kapadia, and D.Metaxas. 2019. Semantic graph convolutional networks for 3d human pose regression. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*

[74] X. Zhou, Q. Huang, X. Sun, X. Xue, and Y. We. 2017. Weakly-Supervised Transfer for 3D Human Pose Estimation in the Wild. In *In International Conference on Computer Vision.*

[75] X. Zhou, X. Sun, W. Zhang, S. Liang, and Y. Wei. 2016. Deep kinematic pose regression. In *In European Conference on Computer Vision.*