Weakly Supervised Generative Network for Multiple 3D Human Pose Hypotheses

Chen Li
lic@comp.nus.edu.sg
Gim Hee Lee
gimhee.lee@comp.nus.edu.sg

Department of Computer Science,
School of Computing,
National University of Singapore

Abstract

3D human pose estimation from a single image is an inverse problem due to the inherent ambiguity of the missing depth. Several previous works addressed the inverse problem by generating multiple hypotheses. However, these works are strongly supervised and require ground truth 2D-to-3D correspondences which can be difficult to obtain. In this paper, we propose a weakly supervised deep generative network to address the inverse problem and circumvent the need for ground truth 2D-to-3D correspondences. To this end, we design our network to model a proposal distribution which we use to approximate the unknown multi-modal target posterior distribution. We achieve the approximation by minimizing the KL divergence between the proposal and target distributions, and this leads to a 2D reprojection error and a prior loss term that can be weakly supervised. Furthermore, we determine the most probable solution as the conditional mode of the samples using the mean-shift algorithm. We evaluate our method on three benchmark datasets – Human3.6M, MPII and MPI-INF-3DHP. Experimental results show that our approach is capable of generating multiple feasible hypotheses and achieves state-of-the-art results compared to existing weakly supervised approaches. Our source code is available at: https://github.com/chaneyddtt/weakly-supervised-3d-pose-generator.

1 Introduction

3D human pose estimation from a monocular image refers to the task of recovering 3D human pose from a 2D image of the person. This task is extensively studied in the computer vision community due to its potentially useful applications in surveillance, healthcare, movie productions, robotics, etc. Most existing works for the task of 3D human pose estimation from a monocular image assume a uni-modal posterior distribution where only a single solution can exist. On the contrary, following the arguments by [14, 18], we reason that 3D human pose estimation from a monocular image is actually an inverse problem with the possibility of multiple feasible solutions due to the inherent ambiguity of the missing depth. Enforcing a uni-modal posterior distribution on the models can lead to overfitting that gives undesirable performance.

To the best of our knowledge, the only existing works that addressed the inverse problem of 3D human pose estimation from a monocular image are Jahangiri and Yullie [14], and Li and Lee [18]. More specifically, [14] uses optimization based method that generates

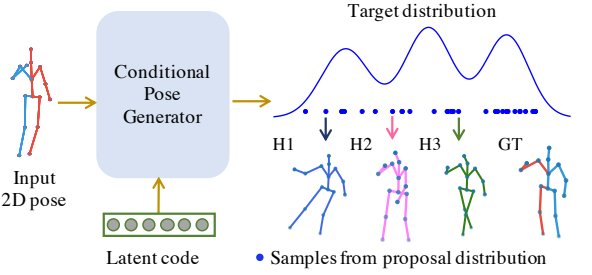

Figure 1: Our deep generative network is conditioned on a input 2D pose. Latent codes are drawn from a normal distribution to generate samples of 3D pose hypotheses that correspond to the target multi-modal posterior distribution.

multiple hypotheses for the inverse problem. Despite the ability to generate multiple hypotheses, the method shows unsatisfactory performance compared to existing deep learning approaches that produce only a single solution. [18] is the first and currently the only deep learning approach that generates multiple hypotheses for the inverse problem of 3D human pose estimation. It uses a mixture density network to model the posterior with a multi-modal mixture-of-Gaussian distribution. Although this approach outperforms other state-of-the-art deep learning single solution approaches, it is supervised that requires a huge amount of ground truth 2D-to-3D correspondences that are often difficult to obtain. To circumvent the need for ground truth data, an increasing number of weakly supervised [10, 15, 29, 30] and unsupervised [26] deep learning approaches are proposed in the recent years. However, these approaches are still based on a uni-modal posterior assumption that gives a single solution to the inverse problem of 3D human pose estimation from a monocular image.

In this paper, we propose a weakly supervised deep generative network to address the inverse problem of 3D human pose estimation. To this end, we design a deep generative network to model a proposal distribution which we use to approximate the unknown multi-modal posterior distribution. Figure 1 shows an illustration of our approach. We achieve the approximation by minimizing the KL divergence between our proposal distribution and the target posterior distribution. This leads to a loss function that minimizes the expectation of a 2D reprojection error and a prior term over the samples drawn from the proposal distribution. The 2D reprojection error ensures that samples of the 3D human pose drawn from our deep generative network reproject closely to the 2D pose observed in the image. We use a discriminator based on the maximum mean discrepancy (MMD) [4, 19] as the prior term to encourage the generated 3D human pose to be "human-like". Furthermore, we prevent the mode collapse problem of our generative network by introducing two additional losses [31, 36] into the prior term to encourage diversity in the generated 3D human poses.

Given an input 2D human pose during inference, we draw samples from the posterior distribution by generating multiple 3D human poses from our generative network. We determine the most probable solution as the conditional mode of the samples using the mean-shift algorithm. We further propose a time-efficient variant to approximate the conditional mode. Specifically, we approximate the conditional mode as the output of our generative network from an all-zero latent code input. Experimental results show that our approach can achieve superior performance compared to state-of-the-art weakly supervised approaches on the Human3.6M dataset [13]. We also test on the MPII [1] and the MPI-INF-3DHP datasets [21] to show the generalization capacity. Our contributions are summarized as: (1) We propose a weakly supervised deep generative network to generate multiple hypotheses for the inverse problem of 3D human pose estimation. (2) We prevent mode collapse of our network by introducing additional losses to encourage diversity of the generated hypotheses. (3) We achieve state-of-the-art results compared to other weakly supervised approaches.

2 Related work

Existing 3D pose estimation approaches can be divided into three categories: Fully and weakly supervised approaches based on a uni-modal posterior, and fully supervised approaches based on a mixture-of-Gaussians distribution.

Most existing works are fully supervised, which train their models either in an end-to-end [17, 21, 24, 28, 35] or a two-stage manner [6, 20, 22, 25, 32]. Pavlakos *et al*. [24] use a volumetric representation for the 3D space and train a deep network to estimate the probability that a joint is located at each voxel. Because of the high dimension of the output, a coarse-to-fine strategy is adopted to finetune the estimation iteratively. To improve the generalization capacity, Zhou *et al*. [35] proposes a transfer learning approach such that the network can be trained with both outdoor and indoor images. For the two-stage approaches, Martines *et al*. [20] use a simple deep neural network to estimate 3D pose from 2D joint detections. Despite the impressive results, these approaches require ground truth 2D-to-3D labels, which are tedious to collect especially for outdoor environments.

More recently, several works begin to focus on weakly supervised [10, 29, 30], unsupervised [26] and self-supervised learning [7]. Wandt *et al*. [30] weakly supervise their network with only 2D ground truth labels by projecting the estimated 3D pose into 2D space. A critic network is then used to enforce the estimated poses to be realistic. Chen *et al*. [7] propose a self-supervised learning framework that lifts the 2D input to 3D pose, projects the 3D pose after a random transform, lifts the projection to 3D, undo the random transform and then projects back onto the 2D image. A self-consistency constraint and a 2D pose discriminator is applied on the original and final 2D poses to enable the lifting network to estimate valid 3D poses. The discriminators applied in both approaches play a key role to enforce valid estimations.

All of the above mentioned approaches assume a uni-modal posterior distribution, and only estimate one 3D pose for each 2D input. Two recent works [14, 18] explore a new line of research in generating multiple hypotheses for 3D human pose estimation. They argue that 3D pose estimation from 2D observations is an inverse problem where multiple solutions exist. To generate the multiple solutions, Jahangiri and Yullie [14] learn an occupancy matrix to represent the plausible angular regions for each joint, and then generate multiple hypotheses by sampling from the occupancy matrix. Li and Lee [18] use a mixture density network (MDN) to learn the multi-modal posterior distribution and take the conditional mean values of the mixture-of-Gaussian distribution as the hypotheses. Although [18] achieves promising results, the method is strongly supervised and require ground truth 2D-to-3D correspondences for training. In contrast to [14, 18], we propose a weakly supervised generative model to generate multiple 3D pose hypotheses.

3 Our Method

We propose a weakly supervised approach to generate multiple hypotheses from a given 2D human pose input. Let $\mathbf{x} \in \mathbb{R}^{2C}$ denotes the 2D joint detection, where C is number of joints in a skeleton. We generate multiple 3D pose hypotheses $\mathbf{y} \in \mathbb{R}^{3C}$ for each 2D pose input, where all the hypotheses reproject close to the 2D pose input. The true posterior $P(\mathbf{y} \mid \mathbf{x})$ is a multi-modal distribution because of the depth ambiguity and occluded joints. We design a deep generative network as a proposal distribution $Q(\mathbf{y} \mid \mathbf{x})$ to approximate the unknown target posterior distribution $P(\mathbf{y} \mid \mathbf{x})$. Figure 2 shows the deep generative network that we designed as the proposal distribution $Q(\mathbf{y} \mid \mathbf{x})$. It consists of four main components: (1) a

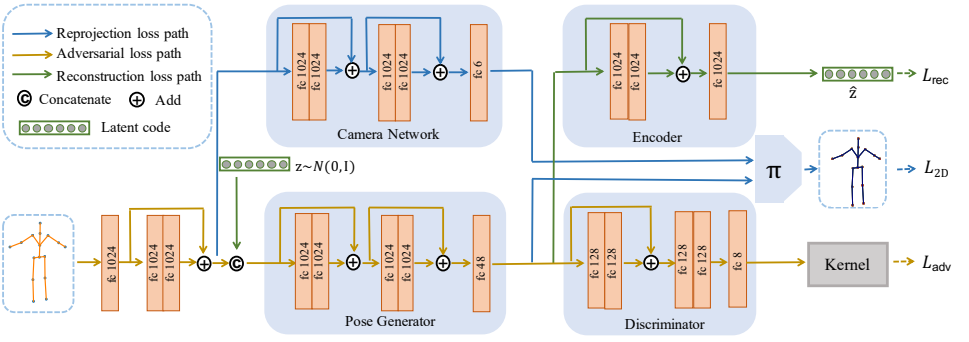

Figure 2: Our deep generative network to generate multiple 3D human pose hypotheses.

pose generator network that generates a 3D pose hypothesis $\mathbf{y}$ from on an input 2D pose $\mathbf{x}$ and latent code $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$; (2) a camera network that estimates the camera matrix $\mathbf{M} \in \mathbb{R}^{2\times 3}$ to project the generated 3D pose hypotheses into the 2D space; (3) a discriminator as the prior $P(\mathbf{y})$ of the generated 3D pose; and (4) an encoder as a second prior to prevent the model collapse of our generative model.

3.1 Conditional Pose Generator

Our goal is to train the model $Q(\mathbf{y} \mid \mathbf{x})$ to generate samples of 3D pose hypotheses from the unknown target posterior distribution $P(\mathbf{y} \mid \mathbf{x})$. To this end, we minimize the KL divergence between the proposal $Q(\mathbf{y} \mid \mathbf{x})$ and the target posterior $P(\mathbf{y} \mid \mathbf{x})$ distributions:

$$\mathcal{L} = KL[Q(\mathbf{y} \mid \mathbf{x}) \| P(\mathbf{y} \mid \mathbf{x})] + H(Q(\mathbf{y} \mid \mathbf{x})). \tag{1}$$

Following [8], we also minimize the entropy of the proposal distribution so that the output 3D pose $\mathbf{y}$ learns enough information from the 2D input $\mathbf{x}$. According to the definition of KL divergence and entropy, the objective function is evaluated as:

$$\mathcal{L} = -\sum_{y} Q(\mathbf{y} \mid \mathbf{x}) \log \frac{P(\mathbf{y} \mid \mathbf{x})}{Q(\mathbf{y} \mid \mathbf{x})} - \sum_{y} Q(\mathbf{y} \mid \mathbf{x}) \log Q(\mathbf{y} \mid \mathbf{x}) = -\sum_{y} \{Q(\mathbf{y} \mid \mathbf{x}) \log P(\mathbf{y} \mid \mathbf{x})\}. \tag{2}$$

We get our final objective function by applying the Bayes rule on $P(\mathbf{y} \mid \mathbf{x}) = \frac{P(\mathbf{x}|\mathbf{y})P(\mathbf{y})}{P(\mathbf{x})}$:

$$\mathcal{L} = -\mathop{\mathbb{E}}_{y\sim Q(\mathbf{y}|\mathbf{x})} \{\log P(\mathbf{x} \mid \mathbf{y}) + \log P(\mathbf{y})\}. \tag{3}$$

The objective function consists of a likelihood term $P(\mathbf{x} \mid \mathbf{y})$ and a prior $P(\mathbf{y})$ term after we drop the constant term $\log P(\mathbf{x})$. We represent the likelihood term $P(\mathbf{x} \mid \mathbf{y})$ by a Laplace distribution:

$$P(\mathbf{x} \mid \mathbf{y}) = \frac{1}{2b} \exp - \frac{|\pi(\mathbf{y}) - \mathbf{x}|}{b}, \tag{4}$$

where b is the scale parameter, $\pi(\mathbf{y})$ and $\mathbf{x}$ are the 2D reprojection of the generated 3D pose and the input 2D pose, respectively. Note that the Gaussian or the Laplacian distribution can be used for the likelihood term, and we chose the Laplacian distribution due to its robustness

to noisy and outlier 2D joint inputs. Inspired by [12, 30], we estimate a camera matrix $\mathbf{M} \in \mathbb{R}^{2\times 3}$ from the 2D observation by using a camera network. The generated 3D pose $\mathbf{y}$ and camera matrix $\mathbf{M}$ are fed into a reprojection module to get the 2D reprojection. Under a weak perspective camera assumption, the 2D reprojection of the generated pose $\mathbf{y}$ is given by $\pi(\mathbf{y}) = \mathbf{M}\mathbf{y}$. Maximizing the log-likelihood term is equivalent to minimizing the reprojection error, which results in our 2D loss: $\mathcal{L}_{2D} = |\mathbf{M}\mathbf{y} - \mathbf{x}|$.

The prior term $P(\mathbf{y})$ represents the prior knowledge of real 3D poses, *e.g.* bone length, joint angle limit and symmetric information, and we use the discriminator from the MMD GAN [4, 19] to learn the prior knowledge from a set of 3D poses. Note that it is not necessary for this set of 3D poses to be the ground truth labels of the respective input 2D poses. The input to the discriminator is a concatenation of the 3D pose and the corresponding KCS matrix [30]. As shown in Figure 2, our pose generator has similar structure to a conditional GAN. The generator generates pose hypotheses from latent code $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ conditioned on the input 2D pose $\mathbf{x}$, while the discriminator try to distinguish the generated poses from real poses. Consequently, a sample $\mathbf{y}$ is drawn from the proposal distribution in Equation (3) as:

$$\mathbf{y} \leftarrow Q(\mathbf{y} \mid \mathbf{x}, \mathbf{z} \sim \mathcal{N}(0, \mathbf{I})). \tag{5}$$

3.2 Diverse Pose Hypotheses

Minimization of the KL divergence between the proposal distribution and the target posterior distribution may result in the generator learning only a subset of the target posterior distribution. This phenomenon known as the mode collapse problem [2, 27] is widely discussed in the GAN literature. This problem manifests itself in the generator generating the same poses for different input latent codes conditioned on the same input 2D pose. To circumvent this problem, we add a second prior with a regularizer [31] to explicitly encourage diversity and an encoder to reconstruct the input noise [36].

Let $G(\mathbf{x}, \mathbf{z}_1)$ and $G(\mathbf{x}, \mathbf{z}_2)$ denote the output of the generator given 2D observation $\mathbf{x}$, latent codes $\mathbf{z}_1$ and $\mathbf{z}_2$ sampling from $\mathcal{N}(0, \mathbf{I})$. We encourage the generator to generate diverse hypotheses by maximizing the objective:

$$\mathcal{L}_{\text{reg}} = \mathbb{E}_{\mathbf{z}_1, \mathbf{z}_2}[\min(\frac{|G(\mathbf{x}, \mathbf{z}_1) - G(\mathbf{x}, \mathbf{z}_2)|}{|\mathbf{z}_1 - \mathbf{z}_2|}, \tau)], \tag{6}$$

where τ is a constant to ensure numerical stability. The regularizer forces the generator to generate diverse poses depending on the distance between the input latent codes.

To further prevent the mode collapse, we also introduce another encoder E to reconstruct the input latent code [36]:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{\mathbf{z} \sim \mathcal{N}(0,1)} |\mathbf{z} - E(G(\mathbf{x}, \mathbf{z}))|. \tag{7}$$

The reconstruction loss encourages the connection between the output 3D pose and input latent code to be invertible, such that it helps prevent the many-to-one mapping problem in mode collapse. Intuitively, if $G(\mathbf{x}, \mathbf{z}_1)$ and $G(\mathbf{x}, \mathbf{z}_2)$ are the same when $\mathbf{z}_1 \neq \mathbf{z}_2$, we can never recover $\mathbf{z}_1$ or $\mathbf{z}_2$ because the inputs to the encoder E are the same.

3.3 Optimization

Inspired by the MMD GAN [4, 19], we use the kernel maximum mean discrepancy to distinguish the generated and real data distributions. The unbiased estimator of the squared MMD

is given by:

$$MMD_u^2(P,Q) = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(\mathbf{y}_i,\mathbf{y}_j) + \frac{1}{n(n-1)}\sum_{i\neq j}^{m} k(\tilde{\mathbf{y}}_i,\tilde{\mathbf{y}}_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(\mathbf{y}_i,\tilde{\mathbf{y}}_j). \quad (8)$$

where $\mathbf{y} \sim P(\mathbf{y})$ and $\tilde{\mathbf{y}} \sim Q(\mathbf{y} \mid \mathbf{x})$ represent samples from the real and generated distributions respectively. We adopt a mixed kernel consisting of the rational quadratic (RQ) kernel and the dot kernel: $k^{rq*} = k^{rq} + k^{dot}$ following [4], where

$$k_\alpha^{rq}(x_1,x_2) = (1 + \frac{\|x_1 - x_2\|^2}{2\alpha})^{-\alpha}, k^{dot}(x_1,x_2) = \langle x_1, x_2 \rangle. \quad (9)$$

The pose generator tries to fool the discriminator by generating realistic poses, hence it minimizes a adversarial loss given by: $\mathcal{L}_{\text{adv}} = MMD_u^2(P,Q)$. At the same time, the pose generated from the same 2D input should be diverse and also keep consistent with the 2D input. Finally, the full objective function of the generator is expressed as:

$$\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{2D}}\mathcal{L}_{\text{2D}} - \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}}, \quad (10)$$

where λ_{2D}, λ_{reg} and λ_{rec} represent the weights of the corresponding losses. On the other hand, the discriminator tries to distinguish the real and fake distributions by minimizing $\mathcal{L}_D = -\mathcal{L}_{\text{adv}} + \lambda_{\text{gp}}\mathcal{L}_{\text{gp}}$. The gradient penalty $\mathcal{L}_{\text{gp}}$ term [9] is added to enforce the Lipschitz constraint. The camera estimation network also optimizes a camera loss [30] such that it fulfils the weak perspective camera constraint.

3.4 Best Pose Selection

After training, the generator can generate 3D pose hypotheses for the same 2D input by sampling latent code $\mathbf{z}$ from $\mathcal{N}(0,\mathbf{I})$. In practice, we also want to find the most probable 3D pose from the multiple hypotheses, *i.e.*, the best conditional mode of the posterior distribution. Let $\mathbf{H} = \{\mathbf{h}_1,\mathbf{h}_2,\mathbf{h}_3,...,\mathbf{h}_N\}$ be the pose hypotheses generated from $\mathbf{Z} = \{\mathbf{z}_1,\mathbf{z}_2,\mathbf{z}_3,...,\mathbf{z}_N\}$ conditioned on $\mathbf{x}$, where N is the number of samples. To find the pose with the highest probability, we employ a local mode-finding approach based on mean-shift [2] with a Gaussian kernel. The kernel density estimator is given by:

$$\hat{f}(\mathbf{h}) = \frac{1}{Nw^d}\sum_{i=1}^{N} K(\frac{\mathbf{h}-\mathbf{h}_i}{w}), \quad (11)$$

where w, d and K represent the bandwidth, feature dimension and kernel function, respectively. However, the mean-shift algorithm is computationally expensive especially when the number of samples is large and might be unsuitable for scenario where efficiency is the priority. We propose an alternative method to improve the efficiency. We directly feed an all-zero code $\mathbf{z}$ into the generator and obtain the final pose. This is similar to the 'zero code' used in [5, 33] to obtain a most likely single view depth. Intuitively, an all-zero code is the most likely code because we sample $\mathbf{z}$ from $\mathcal{N}(0,\mathbf{I})$ during training. We will show in the experiments that the zero code can achieve comparable results with the mean-shift algorithm.

4 Experiments

Implementation Details. We train our model with ADAM optimizer with an initial learning rate of 0.0001 and decay every epoch with a decay rate of 0.94. The weights for different losses λ_{gp}, λ_{2D}, λ_{reg} and λ_{rec} are set to 0.1, 10.0, 7.5 and 10.0 respectively.

Datasets. We evaluate our approach on three 3D human pose estimation benchmarks: Human3.6M [13], MPI-INF-3DHP [21] and MPII datasets[1]. The human3.6M dataset is the largest and most commonly used dataset for 3D human pose estimation. There are 15 daily activities in total performed by 7 professional actors under 4 camera views. The MPI-INF-3DHP is a recently proposed dataset which includes both indoor and outdoor scenes. The MPII dataset a challenging benchmark for 2D human pose estimation because of the complex background and severe occlusion. We train our model on the Human3.6M dataset and show results on all three datasets

Data Preprocessing. Following previous work [30], we align every 3D pose in the Human3.6M dataset to a template by applying a transformation to the 3D pose. The transformation, which includes a scale, rotation and translation, is obtained from procrustes analysis on the hip and shoulder joints. Both 2D and 3D poses are centered at the root joint, and each 2D pose is further normalized by dividing its standard deviation. Following previous work [18, 20], we use the stacked hourglass network [23] trained on both MPII and Human3.6M datasets to estimate 2D poses from images.

Evaluation Protocols. Following the standard protocol for Human3.6M dataset [16], we use subjects 1, 5, 6, 7 and 8 for training, and evaluation is done on every 64^{th} frame of subjects 9 and 11. The evaluation metric is the Mean Per Joint Position Error (MPJPE) measured in millimeters. The 3D Percentage of Correct Keypoints (3DPCK) under 150mm radius [21] is adopted as the metric for the MPI-INF-3DHP dataset.

Protocol #2	MH	WS	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez [16]			39.5	43.2	46.4	47.0	51.0	56.0	41.4	40.6	56.5	69.4	49.2	45.0	49.5	38.0	43.1	47.7
Zhou [32]			29.1	34.9	29.9	32.6	31.2	32.3	27.0	33.3	37.6	45.9	32.2	31.5	34.5	22.9	25.9	32.1
Li [18](BH)	✓		35.5	39.8	41.3	42.3	46.0	48.9	36.9	37.3	51.0	60.6	44.9	40.2	44.1	33.1	36.9	42.6
Tung [29]		✓	77.6	91.4	89.9	88.0	107.3	110.1	75.9	107.5	124.2	137.8	102.2	90.3	78.6	-	-	97.2
Wandt [30]		✓	53.0	58.3	59.6	66.5	72.8	71.0	56.7	69.6	78.3	95.2	66.6	58.5	63.2	57.5	49.9	65.1
Drover [6]		✓	60.2	60.7	59.2	65.1	65.5	63.8	59.4	59.4	69.1	88.0	64.8	60.8	64.9	63.9	65.2	64.6
Ours (ZC)	✓	✓	**42.1**	**44.7**	**45.4**	**51.0**	**49.3**	**51.5**	**41.2**	**46.2**	**57.5**	**70.8**	**48.7**	**44.1**	**50.8**	**42.1**	**43.7**	**48.7**
Ours (MS)	✓	✓	41.4	44.3	44.6	50.2	49.3	51.8	40.1	46.2	57.7	72.7	48.7	45.4	49.6	43.8	43.3	48.7
Ours (BH)	✓	✓	38.5	41.7	39.6	45.2	45.8	46.5	37.8	42.7	52.4	62.9	45.3	40.9	45.3	38.6	38.4	44.3
Ours (GT+BH)	✓	✓	26.8	31.2	26.9	33.0	31.0	36.9	28.7	31.2	36.6	46.4	30.0	30.8	31.5	24.7	27.2	31.6

Table 1: Quantitative results of MPJPE on the Human3.6M dataset under protocol #2. The best results for weakly supervised methods are marked in bold. (Our results under ZC setting is used for fair comparison.)

4.1 Quantitative Results on Human3.6M Dataset

The poses generated by the generator are in the template frame as described in the Data Preprocessing. Consequently, we evaluate the effectiveness of the pose generator by showing the MPJPE under protocol #2, where a rigid alignment is applied to the estimated pose before comparison with the ground truth. Table 1 shows the results of our approach and other state-of-the-art fully and weakly supervised approaches. 'MH' represents approaches that generate multiple hypotheses and 'WS' represents weakly supervised approaches. We evaluate our approach under both mean-shift (MS) and zero code (ZC) settings, where we

Protocol #1	MH	WS	Direct.	Discuss	Eating	Greet	Phone	Photo	Pose	Purch.	Sitting	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez [20]			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
Sun [28]			47.5	47.7	49.5	50.2	51.4	55.8	43.8	46.4	58.9	65.7	49.4	47.8	49.0	38.9	43.8	49.6
Zhou [30]			34.4	42.4	36.6	42.1	38.2	39.8	34.7	40.2	45.6	60.8	39.0	42.6	42.0	29.8	31.7	39.9
Li [18](BH)	✓		43.8	48.6	49.1	49.8	57.6	61.5	45.9	48.3	62.0	73.4	54.8	50.6	56.0	43.4	45.5	52.7
Wandt [30]		✓	77.5	85.2	82.7	93.8	93.9	101.0	82.9	102.6	100.5	125.8	88.0	84.8	72.6	78.8	**79.0**	89.9
Ours (ZC)	✓	✓	**67.9**	**75.5**	**71.8**	**81.8**	**81.4**	**93.7**	**75.2**	**81.3**	**88.8**	**114.1**	**75.9**	**79.1**	**83.3**	**74.3**	**79.0**	**81.1**
Ours (MS)	✓	✓	66.0	74.7	71.1	80.6	81.1	93.0	73.2	83.7	90.0	117.4	75.8	79.3	82.1	74.4	77.8	80.9
Ours (BH)	✓	✓	62.0	69.7	64.3	73.6	75.1	84.8	68.7	75.0	81.2	104.3	70.2	72.0	75.0	67.0	69.0	73.9
Ours (GT+BH)	✓	✓	54.8	61.9	48.6	63.6	55.8	73.7	59.0	61.3	62.2	85.7	52.8	60.2	57.5	51.3	56.8	60.0

Table 2: Quantitative results of MPJPE on the Human3.6M under protocol #1.The best results for weakly supervised methods are marked in bold. (Our results under ZC setting is used for fair comparison.)

generate a single pose with the highest probability w.r.t. the proposal distribution. Following previous works [14, 18] that generate multiple hypotheses, we also evaluate our approach under the best hypothesis (BH) setting, where we select the best of ten hypotheses according to the ground truth. As can be seen from Table 1, our approach achieves comparable performance with our supervised counterpart [18], which also generates multiple hypotheses, and superior results compared to other weakly supervised approaches [10, 29, 30]. The similar performance achieved by MS and ZC demonstrates that the pose with highest probability can be approximated from the zero code. This significantly improves the efficiency because sampling is not needed in ZC setting. Moreover, the performance under ZC (or MS) is close to BH. This shows that the pose hypothesis with the highest probability is close to the ground truth pose among all hypotheses generated by the generator. 'GT' represents results when using ground truth 2D joints as input, which indicates that our performance can be further improved when 2D detections are more accurate.

We evaluate our model under protocol #1, where the generated poses are transformed into the camera frame with a rotation matrix $\mathbf{R}$ computed from the camera network output $\mathbf{M} \in \mathbb{R}^{2\times 3}$. As shown in Table 2, our approach outperforms state-of-the-art weakly supervised approach [30]. Note that our approach performs worse than our supervised counterpart [18] under this setting, which can be attributed two reasons: (1) we do not use the 2D-to-3D correspondences where the 3D poses are already in the camera frame, and (2) we add constraint to the camera estimation network based on a weak perspective camera assumption, which is not true for the Human3.6M dataset.

Following [14, 18], we also evaluate the robustness of the pose generator by testing on scenarios with missing joints. This is common in realistic scenarios when some joints are severely occluded and cannot be detected. During training, one or two missing joints are randomly selected from the the limb joints including l/r wrist, l/r knee, l/r elbow and l/r ankle. We use the ground truth 2D joints as input and set 2D coordinate of missing joints to zeros. The weights for different losses λ_{gp}, λ_{2D}, λ_{reg} and λ_{rec} are set to 0.1, 20.0, 7.5 and 10.0, respectively. We set the weights for missing joints in the 2D loss $\mathcal{L}_{\text{2D}}$ to zeros because missing joints do not provide any information for the training. The results are shown in Table 3 where the numbers of [18, 20, 30] are based on the public available implementation or checkpoints. We can see that our approach outperforms state-of-the-art weakly supervised approach[30], and achieve comparable results with our supervised counterpart [18].

4.2 Ablation Studies

Do $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{rec}}$ prevent model collapse? We compare our model with and without $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{rec}}$ to verify their effectiveness on diversity. We do the evaluation on two metrics: (1)

Algorithm	MH	WS	Direct.	Discuss	Eating	Greet	Phone	Smoke	Pose	Purch.	Sitting	SitD.	Smoke	Wait	WalkD.	Walk	WalkT.	Avg.
Martinez [21]			36.4	42.4	41.2	43.3	44.2	54.2	43.6	39.2	55.0	58.7	45.2	45.6	46.1	38.2	42.1	45.0
Jahangiri [12]	✓		108.6	105.9	105.6	109.0	105.5	109.9	102.0	111.3	119.6	107.8	107.1	111.3	108.4	107.0	110.3	108.6
Li [18]	✓		31.4	38.5	37.1	37.8	40.2	49.0	37.1	35.1	47.8	56.7	40.7	39.5	40.9	31.2	34.7	39.8
Wandt [30]		✓	36.9	42.2	36.5	43.7	41.4	46.7	40.4	42.0	48.7	57.3	42.0	43.4	42.9	38.4	38.4	42.7
Ours	✓	✓	35.4	41.3	33.7	42.3	39.1	47.1	36.2	46.9	46.4	57.7	38.6	43.0	42.0	34.8	37.0	41.2
Martinez [21]			41.9	48.4	47.8	49.9	51.8	63.5	49.6	44.4	64.7	70.5	52.6	53.4	52.2	46.7	50.1	52.5
Jahangiri [12]	✓		125.0	121.8	115.1	124.1	116.9	123.8	116.4	119.6	130.8	120.6	118.4	127.1	125.9	121.6	127.6	122.3
Li [18]	✓		36.7	42.4	41.6	43.6	46.6	57.0	42.7	39.9	57.0	65.8	46.8	45.4	46.5	36.3	41.0	46.0
Wandt [30]		✓	52.2	62.2	48.4	59.5	56.7	70.6	53.9	57.8	61.5	83.5	57.7	58.6	73.9	58.2	62.8	60.8
Ours	✓	✓	50.9	53.9	49.8	54.8	54.7	65.1	49.4	49.3	63.5	76.1	54.5	54.3	59.8	54.8	56.1	56.4

Table 3: Results with one (the first five rows) or two (the last five rows) missing joints.

randomly sample 10 pose hypotheses for the same 2D input and calculate the standard deviation (STD) of each joint coordinate w.r.t. the root joint; (2) use the farthest point sampling (FPS) [11] to sample 5 diverse hypotheses from 100 random samples and compute the standard deviation (STD-FPS). Table 4a shows the MPJEP under best hypothesis(BH) and zero code(MS) settings, STD and STD-FPS of our model with and without $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{rec}}$. We can see that the pose hypotheses generated by our full model is much more diverse than the model without $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{rec}}$. Moreover, the full model achieves lower error shows the advantage of generating diverse hypotheses. We also show the five hypotheses sampled by FPS in Figure 3. We can see that the generated poses have different degree of diversity depending on the input 2D poses. The 2D reprojections of all 3D pose hypotheses (last column) overlap with each other shows that there are multiple solutions for each 2D input.

Model	MPJPE(BH)	MPJPE(ZC)	STD	STD-FPS
Full model	31.6	35.3	77.4	122.3
Without	36.3	37.4	3.6	4.8

(a)

λ_{reg}	7.0	7.5	8.0	9.0	10.0	11.0	12.0
STD	72.0	77.4	81.0	91.5	98.2	108.2	113.7
MPJEP	33.0	31.6	31.9	32.8	34.2	36.1	38.3

(b)

Table 4: (a):Our model with and without $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{rec}}$. (b): The impact of changing the weights λ_{reg} on the diversity and accuracy

How does λ_{reg} affect the diversity and accuracy? We add the $\mathcal{L}_{\text{reg}}$ to explicitly encourage the diversity of the generated 3D poses, here we analyze the impact of changing the corresponding weight λ_{reg}. Table 4b shows the estimation error under BH setting and diversity (STD) when λ_{reg} is set to 7.0, 7.5, 8.0, 9.0, 10.0, 11.0, 12.0 with weights for other losses fixed. We can see that the STD increases when λ_{reg} gets larger, which verifies that the $\mathcal{L}_{\text{reg}}$ helps to increase the diversity. At the same time, the error also becomes large where we impose overly strong constraint on diversity with high λ_{reg}. Consequently, the value of λ_{reg} should be a trade-off between accuracy and diversity.

4.3 Results on MPI-INF-3DHP and MPII datasets

We test the generalization capacity of our approach on the MPI-INF-3DHP and MPII datasets. The MPI-INF-3DHP dataset includes images under three different scenes: indoor images with (GS) and without green screen background (no GS), outdoor images (Outdoor), and the MPII dataset only includes outdoor images. Table 5 shows the quantitative results of our approach under ZC and BH settings for the MPI-INF-3DHP dataset. Our results is slightly worse than [30] under ZC setting but outperforms other approaches

Algorithm	MH	WS	GS	No GS	Outdoor	All PCK
Mehta* [21]			84.1	68.9	59.6	72.5
Li [18]	✓		70.1	68.2	66.6	67.9
Kanazawa [15]		✓	-	-	-	77.1
Wandt [30]		✓	-	-	-	81.8
Ours(ZC)	✓	✓	82.1	81.0	72.3	79.3
Ours(BH)	✓	✓	86.9	86.6	79.3	85.0

Table 5: Results on the MPI-INF-3DHP dataset.

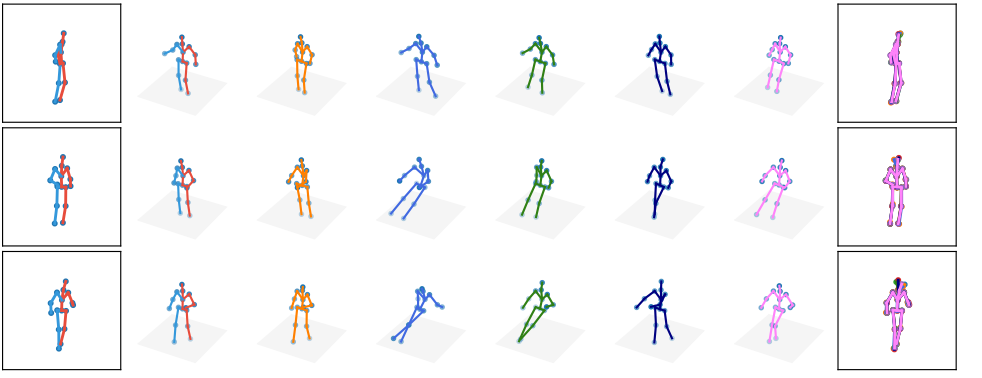

Figure 3: Visualization of five hypotheses sampled by FPS (third to seventh columns). The first and second columns represent the input 2D pose and the corresponding 3D ground truth. The last column shows the 2D reprojections of the five hypotheses (the corresponding 2D reprojection and 3D pose are drawn in the same color).

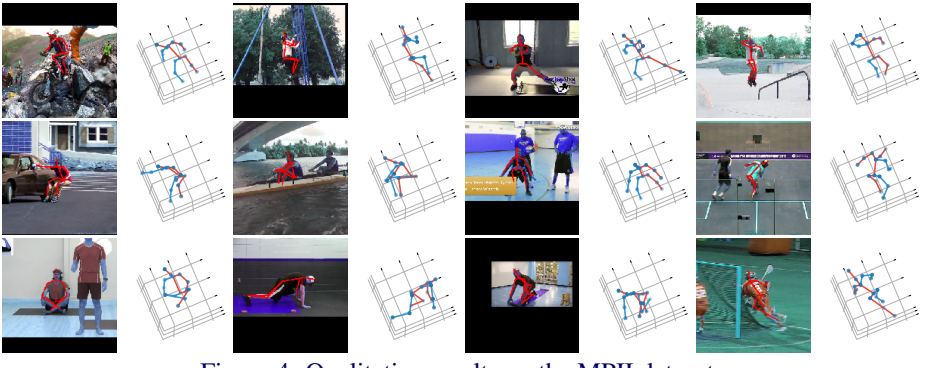

Figure 4: Qualitative results on the MPII dataset.

under BH setting. We only show qualitative results for the MPII dataset because the 3D ground truth is not available. As can been seen from Figure 4, our approach generalizes well to outdoor scenes.

5 Conclusion

We propose a weakly supervised generative network for 3D human pose estimation. Our network is designed to model a proposal distribution and learned by minimizing the KL divergence with the true posterior distribution. Experiments show that our network is able to generate feasible 3D pose hypotheses consistent with 2D reprojections and also achieves better results compared to existing weakly supervised approaches. Moreover, results on the MPII and MPI-INF-3DHP datasets verify the generalization capacity of our network.

References

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

the IEEE Conference on computer Vision and Pattern Recognition, pages 3686–3693, 2014.

[2] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks, 2017.

[3] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223, 2017.

[4] Mikołaj Bińkowski, Dougal J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[5] Michael Bloesch, Jan Czarnowski, Ronald Clark, Stefan Leutenegger, and Andrew J Davison. Codeslam—learning a compact, optimisable representation for dense visual slam. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2560–2568, 2018.

[6] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016.

[7] Ching-Hang Chen, Ambrish Tyagi, Amit Agrawal, Dylan Drover, Stefan Stojanov, and James M Rehg. Unsupervised 3d pose estimation with geometric self-supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5714–5724, 2019.

[8] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.

[9] Dorin Comaniciu and Peter Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):603–619, 2002.

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

[11] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997.

[12] Ikhsanul Habibie, Weipeng Xu, Dushyant Mehta, Gerard Pons-Moll, and Christian Theobalt. In the wild human pose estimation using explicit 2d features and intermediate 3d representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10905–10914, 2019.

[13] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7): 1325–1339, 2013.

[14] Ehsan Jahangiri and Alan L Yuille. Generating multiple diverse hypotheses for human 3d pose consistent with 2d joint detections. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 805–814, 2017.

[15] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7122–7131, 2018.

[16] Muhammed Kocabas, Salih Karagoz, and Emre Akbas. Self-supervised learning of 3d human pose using multi-view geometry. *arXiv preprint arXiv:1903.02330*, 2019.

[17] Kyoungoh Lee, Inwoong Lee, and Sanghoon Lee. Propagating lstm: 3d pose estimation based on joint interdependency. In *European Conference on Computer Vision*, pages 119–135, 2018.

[18] Chen Li and Gim Hee Lee. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.

[20] Julieta Martinez, Rayat Hossain, Javier Romero, and James J Little. A simple yet effective baseline for 3d human pose estimation. In *IEEE International Conference on Computer Vision*, pages 2640–2649, 2017.

[21] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516. IEEE, 2017.

[22] Francesc Moreno-Noguer. 3d human pose estimation from a single image via distance matrix regression. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1561–1570, 2017.

[23] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[24] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3d human pose. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1263–1272, 2017.

[25] Mir Rayat Imtiaz Hossain and James J Little. Exploiting temporal information for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 68–84, 2018.

[26] Helge Rhodin, Mathieu Salzmann, and Pascal Fua. Unsupervised geometry-aware representation for 3d human pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 750–767, 2018.

[27] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

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

[29] Hsiao-Yu Fish Tung, Adam W Harley, William Seto, and Katerina Fragkiadaki. Adversarial inverse graphics networks: Learning 2d-to-3d lifting and image-to-image translation from unpaired supervision. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4364–4372. IEEE, 2017.

[30] Bastian Wandt and Bodo Rosenhahn. Repnet: Weakly supervised training of an adversarial reprojection network for 3d human pose estimation. In *Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[31] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-sensitive conditional generative adversarial networks. *arXiv preprint arXiv:1901.09024*, 2019.

[32] Hashim Yasin, Umar Iqbal, Bjorn Kruger, Andreas Weber, and Juergen Gall. A dual-source approach for 3d pose estimation from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4948–4956, 2016.

[33] Shuaifeng Zhi, Michael Bloesch, Stefan Leutenegger, and Andrew J Davison. Scenecode: Monocular dense semantic reconstruction using learned encoded scene representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11776–11785, 2019.

[34] Kun Zhou, Xiaoguang Han, Nianjuan Jiang, Kui Jia, and Jiangbo Lu. Hemlets pose: Learning part-centric heatmap triplets for accurate 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2344–2353, 2019.

[35] Xingyi Zhou, Qixing Huang, Xiao Sun, Xiangyang Xue, and Yichen Wei. Towards 3d human pose estimation in the wild: a weakly-supervised approach. In *IEEE International Conference on Computer Vision*, pages 398–407, 2017.

[36] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems*, pages 465–476, 2017.