ProPose: Probabilistic 3D Human Pose Estimation with Instance-Level Distribution and Normalizing Flow

Jumin Han, Jun-Hee Kim, Seong-Whan Lee

Department of Artificial Intelligence, Korea University, Seoul, South Korea
{juminhan, jh_kim, sw.lee}@korea.ac.kr

Abstract

3D Human Pose Estimation (HPE) is a one-to-many problem by nature, making it challenging to estimate an accurate 3D pose from a single 2D pose. Some prior works have attempted to tackle this problem by using a conditional generative network. They generate 3D poses from a given 2D pose with noises from a standard Gaussian distribution, while the depth distribution is dependent on each posture and more complex than the standard Gaussian distribution. This may lead to inaccurate distribution learning. In this paper, we propose a probabilistic framework called ProPose to address this issue. ProPose employs Pose Instance-Level Gaussian Distribution (PILGD) derived from 3D pose-based self-representation learning to obtain reliable distribution which is able to address pose-dependent depth distribution. To access this PILGD, we utilize normalizing flow, which learns a mapping function between the PILGD and a 2D Pose-Adaptive Gaussian Distribution (PAGD). This converts the problem of directly estimating 3D poses from 2D poses to a mapping problem between PILGD and PAGD using a normalizing flow. Extensive experiments show the advantages of utilizing the PILGD and PAGD. ProPose achieves comparable performances to previous state-of-the-art probabilistic methods in a multi-hypothesis setting. Notably, ProPose in a single-hypothesis setting demonstrates comparable generalization ability to existing state-of-the-art deterministic methods.

Code — https://github.com/SereneBlooms/ProPose.

Introduction

Lifting-based 3D human pose estimation (HPE) is an essential task in computer vision. It has received considerable attention due to its applications in fields such as AR, VR and robotics (Park and Lee 2011; Zimmermann et al. 2018; Yuan et al. 2021). Recently, researches in this area have made rapid progress with advances in deep neural networks. However, they still suffer from major challenge, which is the inherent problem of depth ambiguity caused by the one-to-many problem (Yang and Lee 2007; Martinez et al. 2017; Kim and Lee 2024).

In recent years, various studies (Martinez et al. 2017; Ci et al. 2019; Pavllo et al. 2019; Zeng et al. 2020) have focused on deterministic 3D HPE and tackled the problems of depth ambiguity. These studies have shown significant improvements in generating plausible 3D poses. However, deterministic method may not be an effective approach under the scenario where there are multiple solutions for a single 2D pose (Li et al. 2022; Li and Lee 2019). In this context, some researches (Sharma et al. 2019; Ci et al. 2023; Oikarinen, Hannah, and Kazerounian 2021; Wehrbein et al. 2021) have focused on multi-hypothesis 3D HPE, which predicts multiple 3D poses for a single 2D pose to deal with depth ambiguity. They have shown substantial potential in multi-hypothesis setting.

Multi-hypothesis 3D HPE generally takes a probabilistic approach to resolve the one-to-many problem. Some of them utilize conditional generative models to generate plausible 3D poses from noises with a given 2D pose (Sharma

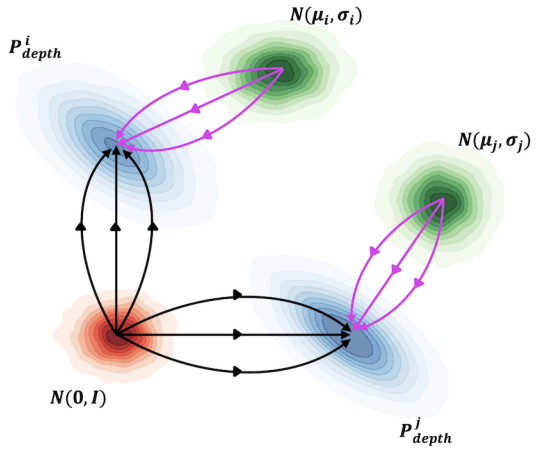

Figure 1: A high level illustrative description of instance-level non-standard Gaussian distribution $N(\mu, \sigma)$ to address depth distribution. P^i_{depth} and P^j_{depth} refer to the depth distributions of two different $pose_i$ and $pose_j$, respectively. The black arrows represent addressing the depth distributions with the same distribution. The pink arrows represent addressing the depth distributions with non-standard Gaussian distributions.

et al. 2019; Ci et al. 2023; Wehrbein et al. 2021). The noises are generally sampled from a standard Gaussian distribution as a prior distribution (Kingma and Welling 2022; Goodfellow et al. 2014; Dinh, Krueger, and Bengio 2015; Ho, Jain, and Abbeel 2020). This process is interpreted as addressing the depth ambiguity of the entire 2D pose with a same distribution, even though there is huge discrepancy between the depth distributions of each 3D pose. For example, the depths of distinctly different poses (*e.g*. sitting and standing) are addressed by the same distribution, which leads to a generation of implausible 3D pose. In addition, the pose-irrelevant prior distribution potentially leads to an inefficient distribution learning without the consideration of the complex dependencies between 2D and 3D poses.

In this paper, we propose probabilistic 3D HPE framework, ProPose, which resolves the pose-dependent depth ambiguity with instance-level non-standard Gaussian distribution as shown in Fig. 1. In designing ProPose, we have two key considerations. **1)** How can we obtain a reliable Gaussian distribution, which is able to address pose-dependent depth distribution? **2)** How can we access this reliable Gaussian distribution? To answer **1)**, we obtain 3D Pose Instance-Level Gaussian Distribution (PILGD) through self-representation learning module based on 3D pose. The module addresses the pose-dependent depth distribution by estimating the Gaussian parameters (mean and variance) for each pose. In this way, the PILGD is defined as an affine transformation of the standard normal distribution. To address **2)**, we utilize normalizing flow to access the PILGD from a base distribution. Rather than using the standard Gaussian distribution as a base distribution, we employ 2D Pose-Adaptive Gaussian Distribution (PAGD) as a base distribution through self-representation learning to obtain more informative base distribution. As a result, we convert the task of directly estimating 3D pose from 2D pose to a task of learning a mapping function between PILGD and PAGD, using normalizing flow.

Extensive experiments comparing to existing state-of-the-art models show that ProPose can generate more plausible 3D poses compared to the other state-of-the-art models. We achieve comparable performances in both multi-hypothesis and single-hypothesis settings. Our contributions are summarized as follows:

- We propose pose instance-level Gaussian distribution to address the pose-dependent depth ambiguity.
- The ProPose is a novel probabilistic approach that converts the 3D HPE to the mapping problem between the two Gaussian distributions.
- To the best of our knowledge, we are the first to utilize self-supervised representations in the field of multi-hypothesis 3D HPE.

Related Work

Our approach belongs to the category of lifting-based 3D HPE, which estimates a 3D pose from a 2D pose estimated by off-the-shelf 2D pose estimator (Newell, Yang, and Deng 2016; Chen et al. 2018; Sun et al. 2019; Fang et al. 2022). Extensive researches have been done on the area of the lifting-based 3D HPE. In this section, we roughly review these methods, starting with deterministic 3D HPE and ending with probabilistic 3D HPE.

Deterministic 3D Human Pose Estimation

This lifting-based 3D HPE is a commonly used approach because it can benefit from the strong performance of 2D pose detectors and is not hindered by background elements in the image. (Park, Hwang, and Kwak 2016) utilized a convolutional neural network to learn the relative 3D joint position information between two different joints. (Moreno-Noguer 2017) formulated the lifting process as a distance matrix regression to estimate the 3D pose. (Martinez et al. 2017) subsequently led to the rapid advancement of lifting-based 3D HPE through the utilization of a basic fully-connected neural network, which enabled the achieving of high performances. However, these methods cannot efficiently address the one-to-many problem because they solved the one-to-many problem with one-to-one problem solver.

Probabilistic 3D Human Pose Estimation

Lifting-based 3D HPE is essentially a one-to-many problem, and resolving depth ambiguity in the lifting process is a challenging issue. To address this, extensive studies (Li and Lee 2019, 2020; Sharma et al. 2019; Wehrbein et al. 2021; Ci et al. 2023) have been done with multi-hypothesis concept to consider the scenario that multiple 3D poses can be matched with a single 2D pose. (Li and Lee 2019) utilized mixture density network to generate multiple hypotheses of 3D pose from a single 2D pose. (Li and Lee 2020) leveraged generative network to generate multiple 3D pose hypotheses from a single 2D pose even in the lack of ground-truth 3D poses. (Sharma et al. 2019) employed conditional variational autoencoder to generate multiple 3D pose hypotheses from a single 2D pose. They especially ranked the generated 3D poses through joint-ordinal depth relations to choose the best 3D pose among the generated hypotheses. More recently, (Wehrbein et al. 2021) tried to solve the one-to-many problem of 3D HPE through normalizing flow. (Ci et al. 2023) leveraged powerful generative network, (Song et al. 2020), to generate multiple 3D poses. Most of these methods generate the 3D pose hypotheses from a given 2D pose and noises sampled from a standard Gaussian distribution. However, this generating process is interpreted as addressing the pose-dependent depth distribution with a same distribution. In addition, the depth distribution is more complex than the standard Gaussian distribution so that they may introduce inefficiency to the training process.

Method

In this section, we describe our probabilistic framework, ProPose. The overall framework and sampling process of ProPose are illustrated in Fig 2. ProPose starts from the base assumption that there exists a reliable 3D Pose Instance-Level Gaussian Distribution (PILGD) that can address individual depth distribution for each human posture. We utilize the self-representation learning to obtain the PILGD. Once we obtain the PILGD, we obtain 2D Pose-Adaptive

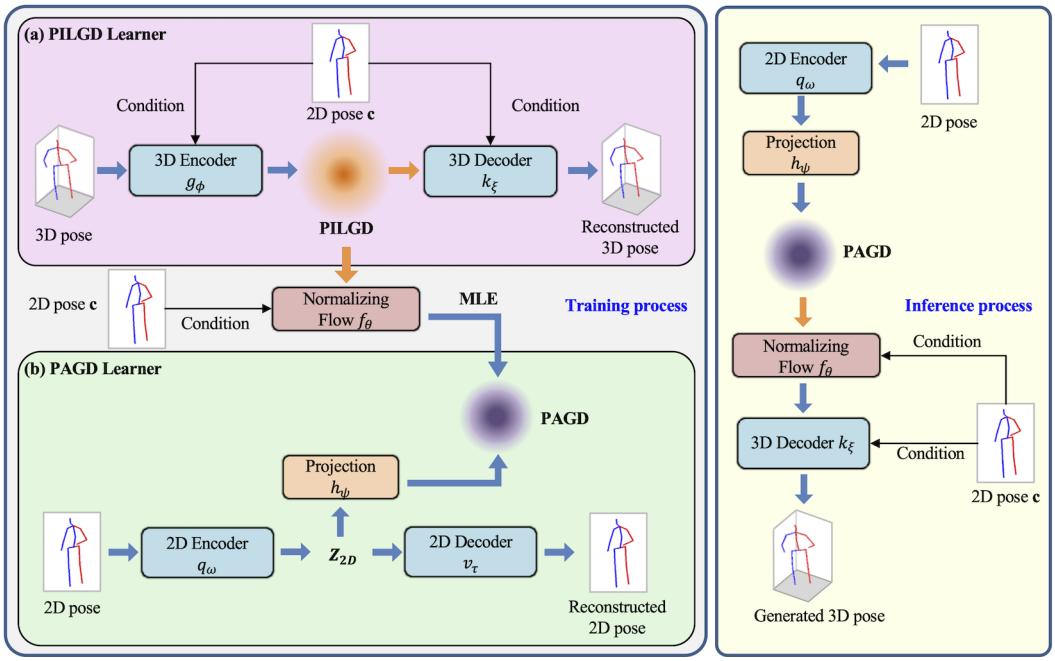

Figure 2: Overall framework of ProPose. The orange arrow refers to sampling from distribution. (a),(b) are illustrations of the self-representation learning processes to obtain PILGD and PAGD, respectively. The normalizing flow maps between the PILGD and PAGD. During inference process, we estimate PAGD from 2D pose and generate 3D pose via the inverse path of normalizing flow and 3D decoder.

Gaussian Distribution (PAGD) through self-representation learning. We then convert the 3D HPE to a learning mapping function between two Gaussian distributions, PILGD and PAGD, using normalizing flow. We describe the detailed method in following sections.

PILGD: Pose Instance-Level Gaussian Distribution

To pursue more accurate generation of 3D pose from 2D pose, we argue that a reliable pose instance-level distribution is needed to address pose-dependent depth ambiguity. We thereby represent a self-representation learning module for 3D poses called PILGD learner, which is based on conditional autoencoder (as shown in Fig. 2a). As a result, for a given deterministic neural encoder $g_\phi : \mathbb{R}^{J\times 5} \mapsto \mathbb{R}^d$ with J and d being the number of body joints and the dimension of latent space of conditional autoencoder, we define the PILGD as a affine transformation of standard normal distribution:

$$S_{3D} = N(g_\phi(P_{3D}, P_{2D}), sI_d), \quad (1)$$
$$S_{3D} = N(\mu_{3D}, sI_d), \quad (2)$$

where $P_{3D} \in \mathbb{R}^{J\times 3}$ and $P_{2D} \in \mathbb{R}^{J\times 2}$ refer to the 3D pose and 2D pose, respectively. The $g_\phi(P_{3D}, P_{2D})$ and s refer to the translation factor and scale factor of $N(0, I_d)$ for addressing the individual depth distribution of each human posture. In order to prevent the variance s from becoming too small during the training process, we set s as a hyperparameter. At the end, for a given decoder $k_\xi : \mathbb{R}^{d+J\times 2} \mapsto \mathbb{R}^{J\times 3}$ of conditional autoencoder, the formulation of PILGD learner with a reparameterization trick can be described as:

$$\hat{P}_{3D} = k_\xi(g_\phi(P_{3D}, P_{2D}) + sz, P_{2D}), \quad (3)$$

where z follows $N(0, I_d)$. We train the learner with Mean Squared Error (MSE) loss function between ground-truth 3D pose and reconstructed 3D pose, which can be described as:

$$L_{\text{recon3D}} = \sum_{i=1}^{J} \|P^i_{3D} - \hat{P}^i_{3D}\|_2, \quad (4)$$

where P^i_{3D} and $\hat{P}^i_{3D}$ represent the i^{th} 3D joint coordinate of the ground-truth 3D pose and reconstructed 3D pose, respectively. In addition, we apply a parameterized soft clamping (Ardizzone et al. 2019) on the top of the g_ϕ to restrict the output of the g_ϕ into the interval $(-\alpha, \alpha)$ for the stability of training PILGD learner. The soft clamping is formulated as:

$$\sigma_\alpha(t) = \frac{2\alpha}{\pi} \arctan \frac{t}{\alpha}. \quad (5)$$

Accessing PILGD Through Normalizing Flow

Now that we have the PILGD, our goal is to accurately access the PILDG. This is achieved through density estimation using normalizing flow. Unlike standard normalizing flow (Dinh, Krueger, and Bengio 2015; Dinh, Sohl-Dickstein, and Bengio 2017; Kingma and Dhariwal 2018), we leverage PAGD as a base distribution instead of standard normal distribution to leverage more informative base distribution.

To obtain PAGD, we represent a self-representation learning module for 2D poses called PAGD learner, which is

based on autoencoder as shown in Fig. 2b. We estimate Gaussian parameters of the PAGD from the self-supervised latent representation via the projection layer. We train the autoencoder with L1 loss function between input 2D pose and reconstructed 2D pose, which can be described as:

$$L_{\text{recon2D}} = \sum_{i=1}^{J} \| P_{2D}^{i} - \hat{P}_{2D}^{i} \|_1, \tag{6}$$

where $P_{2D}^{i}, \hat{P}_{2D}^{i} \in \mathbb{R}^{J\times 2}$ represents the i^{th} joint of the input 2D pose and reconstructed 2D pose, respectively. The process to obtain PAGD can be formulated as:

$$\mu_{2D}, \sigma_{2D} = h_\psi(q_\omega(P_{2D})), \tag{7}$$

$$S_{2D} = N(\mu_{2D}, \sigma_{2D}), \tag{8}$$

where q_ω and h_ψ are the encoder of autoencoder and projection layer, respectively. In this manner, we can obtain PAGD, S_{2D}. Additionally, we applied the parameterized soft clamping (Ardizzone et al. 2019) on top of the h_ψ to ensure training stability.

To utilize normalizing flow as a mapping function between the PILGD and PAGD, we train normalizing flow f_θ with maximum likelihood estimation. The likelihood of the value w sampled from S_{3D} in Eq. 2 can be expressed with PAGD parameters using change of variable formula over factorized distribution, which can be described as:

$$p_W(w; c) = N(f_\theta(w; c); \mu_{2D}, \sigma_{2D}) \left| \det \frac{\partial f_\theta}{\partial w} \right|, \tag{9}$$

where $p_W(\cdot)$ and c refer to likelihood function and 2D pose, respectively. We utilize negative log-likelihood of Eq. 9 as a loss function L_{NLL} to maximize the $p_W(w)$. The L_{NLL} can be formulated as:

$$L_{\text{NLL}} = \log \sigma_{2D} + \frac{1}{2} \times \frac{(f_\theta(w; c) - \mu_{2D})^2}{\sigma_{2D}^2} - \log \left| \det \frac{\partial f_\theta}{\partial w} \right| + \frac{1}{2} \times \log 2\pi. \tag{10}$$

By training the f_θ, we can access the self-supervised PILGD from PAGD via inverse path of f_θ.

Adversarial Training

We apply discriminator to the PILGD learner to prevent the generation of implausible 3D pose. A slightly modified discriminator from energy-based generative model (Zhao, Mathieu, and LeCun 2017) is used for PILGD learner. In detail, we use a conditional autoencoder with condition of 2D pose as discriminator. This allows the discriminator to learn a manifold of plausible 3D poses for a given 2D pose. For a given positive margin m, the loss functions are identical to those presented in (Zhao, Mathieu, and LeCun 2017) and can be described as:

$$\begin{aligned} L_{\text{D}} &= D(P_{3D}, P_{2D}) + \left[m - D(\hat{P}_{3D}, P_{2D}) \right]^+, \\ L_{\text{G}} &= D(\hat{P}_{3D}, P_{2D}), \\ L_{\text{PT}} &= \frac{1}{N_b(N_b-1)} \sum_i \sum_{j\neq i} \left(\frac{S_i^T S_j}{\|S_i\| \|S_j\|} \right)^2, \\ L_{\text{disc}} &= L_{\text{D}} + L_{\text{G}} + L_{\text{PT}}, \end{aligned} \tag{11}$$

Algorithm 1: Training scheme of the ProPose

Input: g_ϕ^0, k_ξ^0, f_θ^0, h_ψ^0, q_ω^0, and v_τ^0 (initial weights of ProPose components). N_{p1} and N_{p2} (the number of batches for each phase). α and β (learning rates).

$g_\phi, k_\xi, f_\theta, h_\psi, q_\omega, v_\tau \leftarrow g_\phi^0, k_\xi^0, f_\theta^0, h_\psi^0, q_\omega^0, v_\tau^0$
while The weights of the ProPose have not converged **do**
  for $t = 1, \ldots, N_{\text{p1}}$ **do**
    Compute L_{recon3D} and L_{Bi}
    $L_{\text{phase1}} = L_{\text{recon3D}} + L_{\text{Bi}}$
    $g_\phi \leftarrow g_\phi - \alpha \nabla_{g_\phi} L_{\text{phase1}}$
    $k_\xi \leftarrow k_\xi - \alpha \nabla_{k_\xi} L_{\text{phase1}}$
  end for
  for $t = 1, \ldots, N_{\text{p2}}$ **do**
    Compute L_{recon2D}, L_{NLL}, and L_{m2m}
    $L_{\text{phase2}} = L_{\text{recon2D}} + L_{\text{NLL}} + L_{\text{m2m}}$
    $q_\omega \leftarrow q_\omega - \alpha \nabla_{q_\omega} L_{\text{phase2}}$
    $v_\tau \leftarrow v_\tau - \alpha \nabla_{v_\tau} L_{\text{phase2}}$
    $h_\psi \leftarrow h_\psi - \alpha \nabla_{h_\psi} L_{\text{phase2}}$
    $f_\theta \leftarrow f_\theta - \beta \nabla_{f_\theta} L_{\text{phase2}}$
  end for
end while

where D represents the discriminator and $[\cdot]^+$ refers to $max(0, \cdot)$. The S is a batch of the outputs of encoder and N_b refers to the batch size. Note that the pulling-away term L_{PT} is applied in the PILGD learner loss, not in the discriminator loss.

Additional Loss Functions

Although the loss functions in the previous sections are sufficient to generate plausible 3D poses, we apply the loss functions that are helpful to the training process.

Bidirectional Loss The decoder of PILGD learner k_ξ needs to be conditioned on both $f_\theta^{-1}(s_{2D})$ and s_{3D} sampled from S_{3D} in Eq. 2. Inspired by (Wehrbein et al. 2021), we incorporate the best-of-M loss L_{BoM} into our work. This is identical to formulating the sampling process of ProPose as a loss function. In addition, we leverage MSE loss between the randomly sampled 3D pose and the corresponding ground-truth 3D pose. These loss functions can be described as:

$$L_{\text{BoM}} = \sum_{i=1}^{J} \| P_{3D}^{i} - \frac{\sum_{\tilde{P}_{3D} \in H_{\text{topk}}} \tilde{P}_{3D}^{i}}{k} \|_2, \tag{12}$$

$$L_{\text{rand}} = \sum_{i=1}^{J} \| P_{3D}^{i} - \breve{P}_{3D}^{i} \|_2, \tag{13}$$

where H_{topk} is a subset consisting of the top k best 3D pose hypotheses selected from the set of L generated 3D poses. $\breve{P}_{3D}$ refers to a randomly generated 3D pose hypothesis. In order to regulate the impact of these losses on k_ξ, we employ a random selection function, which can be described as:

$$L_{\text{Bi}} = \begin{cases} 0 & \text{if } \beta < b \\ L_{\text{BoM}} + L_{\text{rand}} & \text{if } \beta \geq b, \end{cases} \tag{14}$$

Protocol 1	Dir.	Dis.	Eat	Gre.	Phon.	Phot.	Pos.	Pur.	Sit	SitD.	Smo.	Wait	WalkD.	Walk	WalkT.	Avg.
(Martinez et al. 2017) (S = 1)	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
(Li and Lee 2020) (S = 10)	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
(Li and Lee 2019) (S = 5)	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
(Sharma et al. 2019)*	37.8	43.2	43.0	44.3	51.1	57.0	39.7	43.0	56.3	64.0	48.1	45.4	50.4	37.9	39.9	46.8
(Wehrbein et al. 2021)*	38.5	42.5	39.9	41.7	46.5	51.6	39.9	40.8	49.5	56.8	45.3	46.4	46.8	37.8	40.4	44.3
(Ci et al. 2023) (S = 10)†	39.9	44.6	40.2	41.3	46.7	53.6	41.9	40.4	52.1	67.1	45.7	42.9	46.1	36.5	38.0	45.1
(Ci et al. 2023)*†	**31.7**	**35.4**	**31.7**	**32.3**	**36.4**	42.4	**32.7**	**31.5**	**41.2**	52.7	**36.5**	**34.0**	**36.2**	**29.5**	**30.2**	**35.6**
Ours (S = 10)	41.1	44.5	42.5	45.7	47.1	51.7	42.5	46.0	51.9	55.1	49.1	43.3	46.8	38.1	41.2	45.8
Ours*	32.5	36.1	34.1	35.0	37.5	**41.3**	34.0	34.4	41.8	**44.4**	40.1	34.4	37.8	30.4	32.7	36.4
Protocol 2	Dir.	Dis.	Eat	Gre.	Phon.	Phot.	Pos.	Pur.	Sit	SitD.	Smo.	Wait	WalkD.	Walk	WalkT.	Avg.
(Martinez et al. 2017) (S = 1)	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
(Sharma et al. 2019)*	30.6	34.6	35.7	36.4	41.2	43.6	31.8	31.5	46.2	49.7	39.7	35.8	39.6	29.7	32.8	37.3
(Wehrbein et al. 2021)*	27.9	31.4	29.7	30.2	34.9	37.1	27.3	28.2	39.0	43.1	34.2	32.3	33.6	26.1	27.5	32.4
(Ci et al. 2023)*†	26.4	31.5	**27.2**	**27.4**	30.3	36.1	26.8	**26.0**	38.4	45.8	**31.2**	29.2	32.2	**23.1**	**25.8**	30.5
Ours (S = 10)	32.1	35.2	35.8	35.5	36.4	40.4	31.6	34.1	41.5	43.9	39.0	32.1	36.1	29.1	33.0	35.7
Ours*	**25.3**	**28.7**	28.0	27.8	**29.5**	**32.6**	**25.4**	26.5	**34.2**	**36.0**	31.9	**26.1**	**30.2**	23.4	26.2	**28.8**

Table 1: The results on the H3.6M in the MH setting. We evaluate our model under protocol 1 and 2. S refers to the number of generated hypotheses. * represents that S is 200. † means that ground-truth root depth information is used.

Type	Method	Training data	GS	noGS	Outdoor	ALL (PCK↑)	ALL (AUC↑)
Deterministic	(Martinez et al. 2017)	H36M	49.8	42.5	31.2	42.5	17.0
	(Yang et al. 2018)	H36M+MPII	-	-	-	69.0	32.0
	(Ci et al. 2019)	H36M	74.8	70.8	77.3	74.0	36.7
	(Xu and Takano 2021)	H36M	81.5	81.7	75.2	80.1	45.8
	(Zhao, Wang, and Tian 2022)	H36M	80.1	77.9	74.1	79.0	43.8
	(Li et al. 2023)	H36M	**86.2**	84.7	81.9	84.1	**53.7**
Probabilistic	(Ci et al. 2023)*†	H36M	73.5	71.8	80.3	74.5	43.6
	Ours*	H36M	81.3	81.5	80.5	81.2	48.8
	Ours w/o PAGD	H36M	82.5	84.2	81.6	82.9	50.7
	Ours	H36M	83.9	**85.5**	**83.4**	**84.4**	52.1

Table 2: Cross-dataset evaluations on 3DHP. We evaluate our model in terms of PCK and AUC. The probabilistic methods are evaluated in the SH setting. † means that ground-truth root depth information is used. * means that the results when a single hypothesis is randomly sampled.

where the β means a value from $U(0, 1)$ and b is hyper-parameter.

Mode-to-Mode Loss Estimating an accurate and plausible 3D pose in single-hypothesis setting is crucial in probabilistic 3D HPE for practical purposes. The simplest approach to estimate a single-hypothesis 3D pose is to sample the 3D pose from the mode of the PAGD. To ensure that this simple method works well with ProPose, we force the f_θ to map between modes of PILGD and PAGD. We use the Euclidean distance loss between $f_\theta(\mu_{3D})$ and μ_{2D} as the simplest method of enforcing the property of mode-to-mode mapping on f_θ, which can be represented as:

$$L_{\text{m2m}} = \|f_\theta(\mu_{3D}) - \mu_{2D}\|_2. \tag{15}$$

Training Scheme

The PILGD can be roughly viewed as manifold representing plausible 3D poses for a given condition of 2D pose. However, we observe that the end-to-end learning of ProPose yields poor PILGD quality and thereby poor performance of f_θ. Inspired by (Brehmer and Cranmer 2020), we address this issue by separating the training process for PILGD learner from the rest of training processes. In detail, we train the PILGD learner and the rest of ProPose alternately. Thereby we can estimate more accurate PILGD and pursue more accurate density estimation by f_θ. The algorithm for this alternate training scheme is outlined in Alg. 1. We omit the L_{disc} for simplicity.

Experiments

In this section, we describe the dataset and evaluation protocols. We compare the performances in Multi-Hypothesis (MH) and Single-Hypothesis (SH) settings with other state-of-the-art models and show how well ProPose generates plausible 3D poses. We then show the effectiveness of the proposed approach through various experiments.

Type	Method	MPJPE(mm)
Probabilistic	(Li and Lee 2020)	80.9
	(Li and Lee 2019)	62.9
	(Wehrbein et al. 2021)	61.8
	(Ci et al. 2023)†	51.0
	Ours	51.9
Deterministic	(Martinez et al. 2017)	62.9
	(Ci et al. 2019)	52.7
	(Pavllo et al. 2019)	51.8
	(Zhao et al. 2024)‡	51.2
	(Li et al. 2023)	**50.5**

Table 3: The results on the H3.6M under protocol 1 in the SH setting. We report probabilistic and deterministic methods, respectively. † means that ground-truth root depth information is used. ‡ indicates that the results are reported when the image is not used.

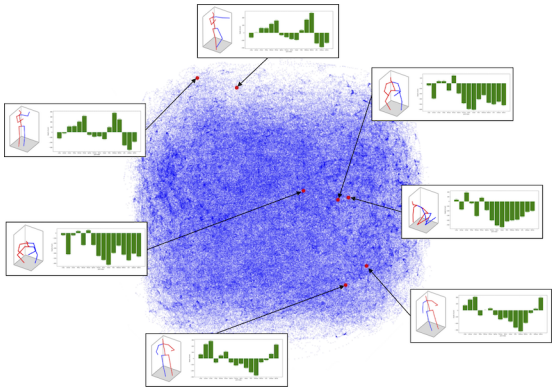

Figure 3: Visualization of the mean value of each PILGD on test set of H3.6M using 2D PCA. Each Black box includes 3D pose (left) and depth distribution of 3D pose (right).

Datasets and Evaluation Metrics

We use two benchmark datasets to evaluate the performance of our model.

Human3.6M (H3.6M) (Ionescu et al. 2013) is a dataset comprises 3D joint positions, 2D projections, and corresponding images, amounting to a total of 3.6 million poses from 11 actors performing 15 sub-activities across 4 camera views. Following convention, subjects 1, 5, 6, 7, and 8 are used for training, while subjects 9 and 11 are used for testing. To measure performance, we employ the Mean Per Joint Position Error (MPJPE) metric following the two standard evaluation protocols. Protocol 1 measures MPJPE between ground-truth and estimated 3D pose after root (hip) centering. Protocol 2 computes MPJPE after applying procrustes alignment. For the MH setting, we calculate MPJPE between ground-truth 3D pose and the best 3D pose hypothesis produced by ProPose. For the SH setting, we generate a single 3D pose deterministically and evaluate under protocol 1 and protocol 2.

MPI-INF-3DHP (3DHP) (Mehta et al. 2017) presents more complex scenarios than H3.6M, including diverse indoor and outdoor settings. For cross-dataset evaluation, we test the model trained on H3.6M directly on 3DHP without finetuning to assess its generalization ability. We report the Percentage of Correctly estimated Keypoints (PCK) at a 150mm threshold and the Area Under Curve (AUC) following convention in the field.

Multi-Hypothesis 3D Pose Estimation on H3.6M

Following the MH setting in the previous works (Wehrbein et al. 2021; Ci et al. 2023; Sharma et al. 2019; Oikarinen, Hannah, and Kazerounian 2021), we generate S 3D pose hypotheses for a given 2D pose and report the MPJPE between ground-truth 3D pose and the best 3D pose hypothesis. Table 1 presents the performance of ProPose with the other state-of-the-art methods in the MH setting. When S is 200, our ProPose achieves comparable performances and even when S is 10, our model shows comparable performances. This indicates that our model generally generate more plausible 3D pose compared to the other methods.

Single-Hypothesis 3D Pose Estimation on H3.6M

We evaluate our ProPose in the SH setting. We generate a single hypothesis from the mode of PAGD. Table 3 presents the performances of ProPose along with other probabilistic methods in a SH setting and deterministic methods. Our probabilistic approach also achieves comparable performances to the deterministic methods.

Cross-Dataset Evaluations

We evaluate ProPose on the 3DHP dataset to assess its generalization ability in the SH setting using ground-truth 2D pose as input following (Martinez et al. 2017; Ci et al. 2019; Pavllo et al. 2019; Li et al. 2023). We report the PCK and AUC values for each method in Table 2. Our model outperforms the other deterministic methods across different scenarios except one scenario for the first time. We also present the performances when a single 3D pose is randomly generated to show how well ProPose addresses depth ambiguity. As we expected, our model outperforms previous state-of-the-art probabilistic method by large margin.

Evaluation of Pose Instance-Level Gaussian Distribution

To demonstrate the discriminative power of PILGD for different human postures, we visualize the mean values μ_{3D} estimated by g_ϕ in Fig. 3. The mean values for 3D poses with different depth distributions are located relatively far from each other, otherwise they are located relatively close. In addition, we report the performances when the PILGD learner is variational autoencoder to show that PILGD is better than standard Gaussian distribution in Table 4. The results demonstrate that the PILGD is more suitable to address the pose-dependent depth distribution.

Effect of 2D Pose-Adaptive Gaussian Distribution

We compare the performances to show that for a base distribution, the PAGD outperforms the standard normal distribu-

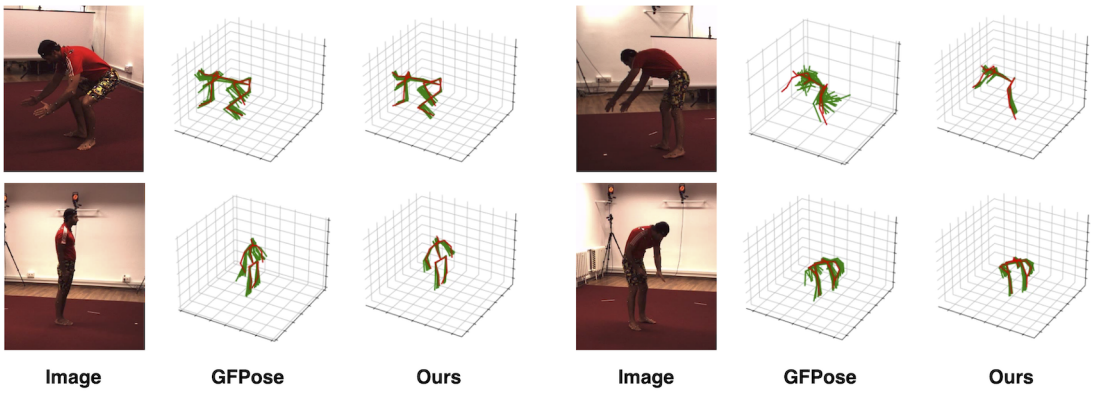

Figure 4: Visualization of generated 3D pose hypotheses (green) from ProPose and ground-truth 3D pose (red).

Type	PILGD	Protocol 1	Protocol 2
MH	$N(\mu_{3D}, \sigma_{3D})$	**36.4**	**28.8**
	$N(0, I)$	39.0	30.8
SH	$N(\mu_{3D}, \sigma_{3D})$	**51.9**	**40.4**
	$N(0, I)$	54.1	41.6

Table 4: We evaluate our model with different PILGD under protocol 1 and protocol 2.

tion. The results for protocol 1 and protocol 2 are reported in Table 5. As we expected, PAGD achieves better performances in the MH setting. Although the results are similar in the SH setting, we achieve better generalization ability when we use the PAGD as the base distribution in the SH setting as shown in Table 2.

Evaluation of 3D Pose Hypotheses

We visualize the generated hypotheses from our model and the state-of-the-art model (Ci et al. 2023), after aligning the 3D pose with root joint in Fig. 4. As we expected, our model generates more plausible and consistent 3D pose hypotheses.

Type	Base distribution	Protocol 1	Protocol 2
MH	$N(\mu_{2D}, \sigma_{2D})$	**36.4**	**28.8**
	$N(0, I)$	37.6	29.7
SH	$N(\mu_{2D}, \sigma_{2D})$	51.9	**40.4**
	$N(0, I)$	**51.8**	40.4

Table 5: We evaluate our model with different base distributions under protocol 1 and protocol 2.

Effect of Training Strategy

To investigate the effectiveness of the proposed training strategy, we report the results for both alternate and end-to-end training schemes. Table 6 shows that the alternate training scheme outperforms the end-to-end training scheme.

Type	Training scheme	Protocol 1	Protocol 2
MH	Alternate	**36.4**	**28.8**
	End-to-End	44.0	35.0
SH	Alternate	**51.9**	**40.4**
	End-to-End	53.2	41.2

Table 6: The results with different training scheme under protocol 1 and protocol 2.

Ablation Study

We have abstracted the additional losses to show its effects. Table 7 shows that each of these losses contributes to the model performances. Note that when we train ProPose without L_{NLL}, we train ProPose in end-to-end training scheme.

Method	Protocol 1	Protocol 2
w/o L_{m2m}	37.6	30.1
w/o L_{Bi}	36.8	29.1
w/o L_{NLL}	41.1	33.0
Ours (Full)	**36.4**	**28.8**

Table 7: Ablation studies on H3.6M in the MH setting.

Conclusions

In this paper, we proposed a novel approach called ProPose with self-representation learning. The key idea of our approach is to build a reliable Gaussian distribution to address the depth ambiguity for each 3D pose in the 2D-to-3D lifting process. Thereby, we convert the 3D HPE problem to distribution mapping problem between two Gaussian distributions by utilizing normalizing flow and pose-adaptive base distribution. Consequently, we can achieve better generalization ability compared to the other methods. Experiments show that our ProPose achieves comparable performances to the state-of-the-art methods and generally generates more plausible 3D poses.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program (Korea University), No. RS-2024-00457882, AI Research Hub Project, and No. RS-2022-II220984, Development of Artificial Intelligence Technology for Personalized Plug-and-Play Explanation and Verification of Explanation).

References

Ardizzone, L.; Lüth, C.; Kruse, J.; Rother, C.; and Köthe, U. 2019. Guided image generation with conditional invertible neural networks. arXiv:1907.02392.

Brehmer, J.; and Cranmer, K. 2020. Flows for simultaneous manifold learning and density estimation. *Advances in Neural Information Processing Systems*, 33.

Chen, Y.; Wang, Z.; Peng, Y.; Zhang, Z.; Yu, G.; and Sun, J. 2018. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7103–7112.

Ci, H.; Wang, C.; Ma, X.; and Wang, Y. 2019. Optimizing network structure for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2262–2271.

Ci, H.; Wu, M.; Zhu, W.; Ma, X.; Dong, H.; Zhong, F.; and Wang, Y. 2023. Gfpose: Learning 3d human pose prior with gradient fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4800–4810.

Dinh, L.; Krueger, D.; and Bengio, Y. 2015. Nice: Non-linear independent components estimation. arXiv:1410.8516.

Dinh, L.; Sohl-Dickstein, J.; and Bengio, S. 2017. Density estimation using real nvp. arXiv:1605.08803.

Fang, H.-S.; Li, J.; Tang, H.; Xu, C.; Zhu, H.; Xiu, Y.; Li, Y.-L.; and Lu, C. 2022. Alphapose: Whole-body regional multi-person pose estimation and tracking in real-time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6): 7157–7173.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2013. 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.

Kim, J.-H.; and Lee, S.-W. 2024. Toward Approaches to Scalability in 3D Human Pose Estimation. *Advances in Neural Information Processing Systems*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 31.

Kingma, D. P.; and Welling, M. 2022. Auto-encoding variational bayes. arXiv:1312.6114.

Li, C.; and Lee, G. H. 2019. Generating multiple hypotheses for 3d human pose estimation with mixture density network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 9887–9895.

Li, C.; and Lee, G. H. 2020. Weakly supervised generative network for multiple 3d human pose hypotheses. arXiv:2008.05770.

Li, H.; Shi, B.; Dai, W.; Zheng, H.; Wang, B.; Sun, Y.; Guo, M.; Li, C.; Zou, J.; and Xiong, H. 2023. Pose-oriented transformer with uncertainty-guided refinement for 2d-to-3d human pose estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1296–1304.

Li, W.; Liu, H.; Tang, H.; Wang, P.; and Van Gool, L. 2022. Mhformer: Multi-hypothesis transformer for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 13147–13156.

Martinez, J.; Hossain, R.; Romero, J.; and Little, J. J. 2017. A simple yet effective baseline for 3d human pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2640–2649.

Mehta, D.; Rhodin, H.; Casas, D.; Fua, P.; Sotnychenko, O.; Xu, W.; and Theobalt, C. 2017. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *Proceedings of the IEEE International Conference on 3D Vision*, 506–516.

Moreno-Noguer, F. 2017. 3d human pose estimation from a single image via distance matrix regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2823–2832.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, 483–499.

Oikarinen, T.; Hannah, D.; and Kazerounian, S. 2021. GraphMDN: Leveraging graph structure and deep learning to solve inverse problems. In *Proceedings of the IEEE International Joint Conference on Neural Networks*, 1–9.

Park, C.-B.; and Lee, S.-W. 2011. Real-time 3D pointing gesture recognition for mobile robots with cascade HMM and particle filter. *Image and Vision Computing*, 29(1): 51–63.

Park, S.; Hwang, J.; and Kwak, N. 2016. 3d human pose estimation using convolutional neural networks with 2d pose information. In *Proceedings of the European Conference on Computer Vision*, 156–169.

Pavllo, D.; Feichtenhofer, C.; Grangier, D.; and Auli, M. 2019. 3d human pose estimation in video with temporal convolutions and semi-supervised training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7753–7762.

Sharma, S.; Varigonda, P. T.; Bindal, P.; Sharma, A.; and Jain, A. 2019. Monocular 3d human pose estimation by

generation and ordinal ranking. In *Proceedings of the IEEE International Conference on Computer Vision*, 2325–2334.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations.*

Sun, K.; Xiao, B.; Liu, D.; and Wang, J. 2019. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5693–5703.

Wehrbein, T.; Rudolph, M.; Rosenhahn, B.; and Wandt, B. 2021. Probabilistic monocular 3d human pose estimation with normalizing flows. In *Proceedings of the IEEE International Conference on Computer Vision*, 11199–11208.

Xu, T.; and Takano, W. 2021. Graph stacked hourglass networks for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16105–16114.

Yang, H.-D.; and Lee, S.-W. 2007. Reconstruction of 3D human body pose from stereo image sequences based on top-down learning. *Pattern Recognition*, 40(11): 3120–3131.

Yang, W.; Ouyang, W.; Wang, X.; Ren, J.; Li, H.; and Wang, X. 2018. 3d human pose estimation in the wild by adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5255–5264.

Yuan, Y.; Wei, S.-E.; Simon, T.; Kitani, K.; and Saragih, J. 2021. Simpoe: Simulated character control for 3d human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7159–7169.

Zeng, A.; Sun, X.; Huang, F.; Liu, M.; Xu, Q.; and Lin, S. 2020. Srnet: Improving generalization in 3d human pose estimation with a split-and-recombine approach. In *Proceedings of the European Conference on Computer Vision*, 507–523.

Zhao, J.; Mathieu, M.; and LeCun, Y. 2017. Energy-based generative adversarial network. arXiv:1609.03126.

Zhao, Q.; Zheng, C.; Liu, M.; and Chen, C. 2024. A single 2d pose with context is worth hundreds for 3d human pose estimation. *Advances in Neural Information Processing Systems*, 36.

Zhao, W.; Wang, W.; and Tian, Y. 2022. Graformer: Graph-oriented transformer for 3d pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20438–20447.

Zimmermann, C.; Welschehold, T.; Dornhege, C.; Burgard, W.; and Brox, T. 2018. 3d human pose estimation in rgbd images for robotic task learning. In *IEEE International Conference on Robotics and Automation*, 1986–1992.