PoseGen: Learning to Generate 3D Human Pose Dataset with NeRF

Mohsen Gholami, Rabab Ward, Z. Jane Wang

University of British Columbia, Vancouver, Canada
{mgholami, rababw, zjanew}@ece.ubc.ca

Abstract

This paper proposes an end-to-end framework for generating 3D human pose datasets using Neural Radiance Fields (NeRF). Public datasets generally have limited diversity in terms of human poses and camera viewpoints, largely due to the resource-intensive nature of collecting 3D human pose data. As a result, pose estimators trained on public datasets significantly underperform when applied to unseen out-of-distribution samples. Previous works proposed augmenting public datasets by generating 2D-3D pose pairs or rendering a large amount of random data. Such approaches either overlook image rendering or result in suboptimal datasets for pre-trained models. Here we propose *PoseGen*, which learns to generate a dataset (human 3D poses and images) with a feedback loss from a given pre-trained pose estimator. In contrast to prior art, our generated data is optimized to improve the robustness of the pre-trained model. The objective of *PoseGen* is to learn a distribution of data that maximizes the prediction error of a given pre-trained model. As the learned data distribution contains OOD samples of the pre-trained model, sampling data from such a distribution for further fine-tuning a pre-trained model improves the generalizability of the model. This is the first work that proposes NeRFs for 3D human data generation. NeRFs are data-driven and do not require 3D scans of humans. Therefore, using NeRF for data generation is a new direction for convenient user-specific data generation. Our extensive experiments show that the proposed *PoseGen* improves two baseline models (SPIN and HybrIK) on four datasets with an average 6% relative improvement. Code is available at https://github.com/mgholamikn/PoseGen.

Introduction

3D human pose and mesh estimation, the task of reconstructing human pose in 3D space given a 2D image of the person, is an ill-posed problem, and many data-driven approaches using deep learning were recently proposed (Liu, Kortylewski, and Yuille 2023; Gholami et al. 2022a; Li et al. 2022). A dataset covering the distribution of possible human poses, global orientation, appearance, and other attributes would be large and difficult to capture in practice. In contrast to vision tasks such as classification and object detection, 3D human pose labels can not be obtained by manual annotation and require an expensive setting for accurate measuring. Therefore, there are limited public datasets and these datasets generally have limited diversity. Unfortunately, most pose estimation models that are trained on public datasets underperform when applied to the out-of-distribution (OOD) samples or samples in the tails of the distribution. Figure 1 shows the domain gap between training and test data.

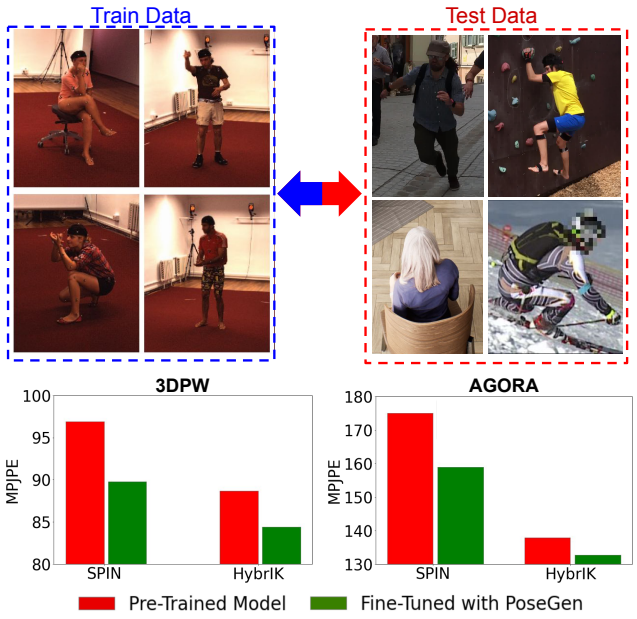

Figure 1: There is a domain gap between the training data used for training pose estimators and in-the-wild images (OOD samples). Therefore, pre-trained models underperform when applied to OOD samples.

To address the above-mentioned concern, prior work proposed augmenting the publicly available datasets by generating 2D-3D human pose pairs (Gong, Zhang, and Feng 2021; Gholami et al. 2022b; Li et al. 2020; Li and Pun 2023) or by rendering synthetic images in more variant poses and appearances (Patel et al. 2021; Black et al. 2023). Augmenting a 3D dataset in the 2D-3D pose space is useful merely for the models that accept 2D poses as input. Therefore, the

majority of 3D pose estimators that accept RGB images as input (not 2D pose) can not be fine-tuned with augmented 2D-3D pose pairs. AGORA (Patel et al. 2021) and BEDLAM (Black et al. 2023) generate synthetic human images and 3D poses. Their experiments show that fine-tuning pre-trained pose estimators on synthetic data makes the models robust on OOD test samples. However, AGORA and BEDLAM use game engines to render human images in an offline manner. Therefore the generated datasets are not optimized for a particular pre-trained model to improve its robustness and generalizability.

Most of the prior arts consider data generation and model training as two different steps (offline methods). PoseAug (Gong, Zhang, and Feng 2021) and AdaptPose (Gholami et al. 2022b) proposed data generation and model training in a single step (online methods). They used feedback from a pose estimator to guide data generation. Online methods make data generation a learnable procedure and prevent generating samples that deviate the downstream model from its objective. PoseAug and AdaptPose use the fixed-hard ratio loss that controls the hardness of generated data. The fixed hard ratio loss converges to zero by either minimizing the loss of the model on the source data or maximizing the loss on the generated data. Such a loss function might converge to generating in-distribution (IND) data. Moreover, fixed hard ratio loss is not source-dataset-free and inevitably needs a source dataset during data generation. Here we propose directly maximizing the loss of a pre-trained model while generating data to find a distribution of data that is OOD for the pre-trained model.

In this work, we leverage the advances in Neural Radiance Fields (NeRF). NeRF can generate high-quality images of the scene from novel views. Compared with traditional rendering engines, NeRF has two major merits: 1) It is differentiable, and 2) it only requires multi-view images for training, and does not require hand-crafted 3D models. Recent works have trained NeRFs on human images and allow the rendering of human images from novel camera viewpoints and novel poses. NeRFs can be trained on user-specific images and therefore can be used to generate user-specific 3D pose datasets. Having such a dataset significantly improves the accuracy of pre-trained models on tasks that require accurate pose estimation (e.g., medical applications).

Figure 2 shows the overall framework of the proposed *PoseGen*. *PoseGen* has a generator that outputs 3D human pose and camera viewpoint. The generated 3D poses are fed to a discriminator that enforces them to be plausible. The generated 3D poses and camera viewpoints are fed to a NeRF model to render corresponding human images. The rendered images are then used to estimate the generated 3D poses. The error of 3D pose estimation is used as feedback to the generator. We investigate two scenarios where we *maximize* or *minimize* the feedback loss from the pose estimator. Maximizing the loss of pose estimator during data generation leads to OOD data generation and minimizing the loss of pose estimator during the data generation results in IND data generation.

In summary, our **contributions** are as follows: We

- propose an end-to-end framework for generating novel user-specific 3D human pose and image datasets.
- propose a generative model that learns the distribution of a pre-trained model and can generate in-distribution and out-of-distribution poses and images.
- propose a simple yet effective feedback function for generative models from pre-trained pose estimators.
- show the effectiveness of NeRF for generating human synthetic datasets.
- obtain SOTA results when doing extensive experiments on 4 datasets with two baseline models.

Related Work

Synthetic 2D-3D Pose Generator. Some prior arts use a two-step method for 3D human pose estimation; In the first step, 2D poses are estimated, and the 3D pose model is trained to estimate 3D from 2D. (Li et al. 2020; Gong, Zhang, and Feng 2021; Gholami et al. 2022b) propose augmenting 2D-3D pose pairs to improve the robustness of pose estimators that use the two-step method. (Li et al. 2020) uses an evolutionary method to augment 3D poses by substituting body parts of real poses from a public dataset. Their evolutionary method is a random data augmentation without any feedback from the pose estimators. (Gong, Zhang, and Feng 2021; Gholami et al. 2022b) propose a learnable framework that learns how to augment data given feedback loss function from the pose estimator. These methods are effective when accurate 2D poses are available at the test time. However, the improvements are limited when 2D pose inputs are in-accurate (Gholami et al. 2022b).

Synthetic Image-3D Generator. Prior works use traditional rendering engines to render photo-realistic human images given human poses. AGORA (Patel et al. 2021) uses 4240 high-quality textured scans of people and randomly samples 3D people and place them in scenes at random distances and orientations. AGORA uses a game engine (Epic Games 2017) optimized for high-quality output to render human images. BEDLAM (Black et al. 2023) uses 271 body shapes with 100 skin textures and 27 different types of hair to the head of SMPL-X. Both AGORA and BEDLAM perform data generation in an offline manner and are unable to generate user-specific datasets. SURREAL applies primitive textures on naked SMPL body mesh to generate synthetic images. SURREAL (Varol et al. 2017) uses 3D sequences of MoCap data and therefore has small variations in terms of body poses.

There is another direction that uses realistic human images and then renders them synthetically in new scenes (Gabeur et al. 2019; Mehta et al. 2017, 2018). These methods have the limitation of real data, including a limited variation of human poses. Based on realistic human images, (Rogez and Schmid 2016) uses 3D pose to select real image whose 2D pose locally matches the projected 3D pose. Selected images are then stitched to generate a new synthetic image. The generated images by these methods are prone to be unrealistic. Since rendering photo-realistic images is challenging, some prior work rendered SMPL mesh as a silhouette or body segments and estimated 3D human poses

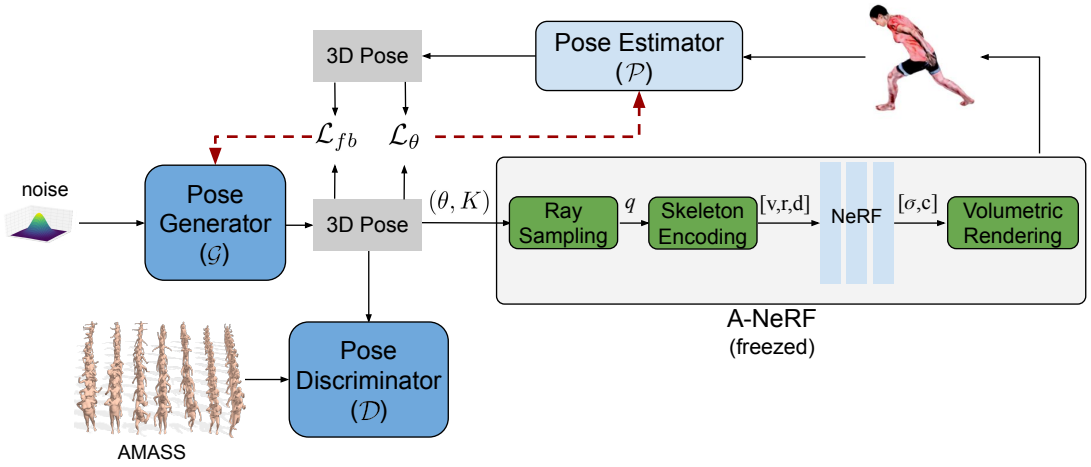

Figure 2: The overall framework of the proposed *PoseGen*. The pose generator learns to generate input of novel poses θ and camera viewpoints K which are fed to a NeRF model to render human images. The 3D pose estimator is trained with rendered images and provides feedback to the pose generator. The feedback function enforces the generator to generate OOD data.

(Xu, Zhu, and Tung 2019; Rong et al. 2019; Pavlakos et al. 2018). These methods do not tackle the performance drops due to unseen human appearances and textures.

Pose Priors. Our work is related to a line of research that uses a machine-learning model to learn the priors of human 3D poses. VPoser (Pavlakos et al. 2019) proposes a Variational Autoencoder to learn a low dimensional latent space for human 3D poses. The learned priors by VAEs are mean-centered and therefore discard the tails of distribution that are far away from the center of the distribution. On the other hand, since its Gaussian prior is unbounded, it is possible to sample poses that are very far from the mean of the distribution which leads to implausible data. (Davydov et al. 2022) proposes adversarial training to learn bounded priors that are able to sample poses far away from the center of the distribution. None of these works is able to learn a distribution of poses that are plausible while OOD for a given pre-trained model. Therefore, these previous works are not able to effectively improve the generalizability of pre-trained models.

Method

Problem Formulation

The overall framework of *PoseGen* is given in Fig 2. It includes a pose generator $\mathcal{G}$ that outputs human poses θ and camera viewpoints K, a Discriminator $\mathcal{D}$ that enforces generated poses to be plausible, a NeRF model that renders human images given poses and camera viewpoints, and a pre-trained 3D pose estimator $\mathcal{P}$ that learns the new data and provides feedback to the generator. The overall objective is to make the 3D pose estimator $\mathcal{P}$ generalizable to unseen (OOD) samples.

The generator samples vector z from a distribution $\mathbb{P}_z \in \mathbb{R}^D$. The output of generator is SMPL body poses $\theta \in \mathbb{R}^{69}$ and camera viewpoint $K \in \mathbb{R}^3$. θ is the relative rotation of limbs in a format of the axis-angle rotation matrix and K is the camera viewpoint in an axis-angle rotation format. The body shape parameters of SMPL β are kept fixed due to the constraints of A-NeRF in rendering bodies with different shapes. Since we aim to generate samples that improve the generalizability of $\mathcal{P}$, it is critical to learn a distribution of data that includes OOD samples. However, generating OOD samples might make the model deviate from the ideal performance. Therefore, we perform two sets of experiments. *Scenario 1:* $\mathbb{P}_z$ is learned to be IND for $\mathcal{P}$, and in *Scenario 2:* $\mathbb{P}_z$ is learned to be OOD for $\mathcal{P}$.

Our method learns the latent space distribution of a generative network $\mathcal{G}$ that outputs the input parameters of a NeRF model $C_\phi(\theta, K)$. In *Scenario 1*, the objective of $\mathcal{G}$ is to generate plausible poses while minimizing the training loss of the pose estimator $\mathcal{P}$:

$$\min_{\mathcal{G},\mathcal{P}} \max_{\mathcal{D}} \mathcal{L}(\mathcal{G}, \mathcal{D}, \mathcal{P}). \tag{1}$$

The learned latent distribution $\mathbb{P}$ can be used to generate human data (images and poses) that will be considered as IND for $\mathcal{P}$. In *Scenario 2* the objective of $\mathcal{G}$ is to generate plausible poses while increasing the training loss of the pose estimator $\mathcal{P}$:

$$\min_{\mathcal{G}} \max_{\mathcal{D},\mathcal{P}} L(\mathcal{G}, \mathcal{D}, \mathcal{P}). \tag{2}$$

The learned latent distribution $\mathbb{P}$ can be used to generate OOD samples for $\mathcal{P}$. The OOD sample contains a distribution of data that covers failure cases of $\mathcal{P}$ including OOD camera viewpoints and human poses. In the following, we will discuss the formulation of $\mathcal{C}, \mathcal{G}$, and $\mathcal{D}$.

Pose Generator and Discriminator

The pose generator samples vector z from a distribution $\mathbb{P}_z$. $\mathbb{P}_z$ is the prior of the synthetic data that plays a crucial role in the quality of generated data. We try different latent space

distributions including normal, uniform, and spherical distribution $\mathcal{S}$, for data generation:

$$z_N \sim \mathbb{P}_z = \mathcal{N}(0,1) \subset \mathbb{R}^d, \tag{3}$$

$$z_U \sim \mathbb{P}_z = \mathcal{U}_{[-1,1]^d} \subset \mathbb{R}^d, \tag{4}$$

$$z_S \sim \mathbb{P}_z = \mathcal{S} \subset \mathbb{R}^d, \tag{5}$$

where the spherical distribution samples vector z_N from normal distribution and computes $z_S = \frac{z_N}{||z_N||_2}$. Prior works suggest that using uniform-like distributions ($\mathcal{U}$ and $\mathcal{S}$) is superior to a normal distribution ($\mathcal{N}$) in learning a general prior for human poses (Davydov et al. 2022). In this work, the generator is not intended to acquire a general prior pose knowledge, but rather to understand the failure modes of the pose estimator $\mathcal{P}$. Uniform-like priors tend to uniformly sample from the plausible poses while the normal distribution is mean-centered and can better find specific modes of the data. Therefore, we argue that normal distribution is a better case for our objective.

In *Scenario 2*, the generator aims to minimize the loss of $\mathcal{P}$ on the generated data. The feedback in this scenario is defined as:

$$\mathcal{L}_{fb} = \frac{1}{N}\frac{1}{J}\sum_{j=1}^{N}\sum_{i=1}^{J}||X_{i,j} - \hat{X}_{i,j}||_2, \tag{6}$$

where X represents the ground-truth 3D poses, $\hat{X}$ denotes the estimated 3D poses, J is the total number of joints, and N is the number of samples. In *Scenario 1*, the generator strives to maximize the loss of $\mathcal{P}$ on the generated data until it reaches a certain threshold, denoted by c. Consequently, the feedback in this scenario is given by:

$$\mathcal{L}_{fb} = c - \frac{1}{N}\frac{1}{J}\sum_{j=1}^{N}\sum_{i=1}^{J}||X_{i,j} - \hat{X}_{i,j}||_2. \tag{7}$$

The overall objective function of the generator is to minimize the weighted summation of adversarial loss and feedback loss. In *scenario 2*, the feedback loss enforces the generator to explore failure modes of $\mathcal{P}$ while the adversarial loss takes care of generated poses being plausible. On the other hand, in *scenario 1* the generator tries to generate novel poses that do not significantly deviate from the original distribution of source data. The overall loss of $\mathcal{G}$ is

$$\mathcal{L}_{\mathcal{G}} = w_1\mathcal{L}_{adv} + w_2\mathcal{L}_{fb}, \tag{8}$$

where $\mathcal{L}_{adv}$ represents the least square GAN loss used for training the generator:

$$\mathcal{L}_{adv} = \mathbb{E}_{z\sim\mathbb{P}_z}[(\mathcal{D}(\mathcal{G}(z)) - 1)^2]. \tag{9}$$

The pose discriminator splits the human body into 6 parts including the torso, left/right leg, and left/right arm and head. The discriminator tries to distinguish real 3D poses from AMASS and synthetic 3D poses from the generator by taking into account the 6 body parts as well as the whole body parts. We use axis-angle joint angles θ as input for the discriminator $\mathcal{D}$. AMASS includes archives of human poses and we assume that using AMASS as the prior for the discriminator does not enforce the generated data to a specific sub-mode of human poses. In the appendix, we show the distribution of AMASS and the distribution of body poses in publicly available datasets such as 3DPWS (test-set). AMASS covers all models of the 3DPW test set and qualitatively proves that using AMASS is not problematic.

The discriminator $\mathcal{D}$ only enforces generated data in terms of body poses θ and does not take into account camera viewpoint K. The camera viewpoint is mainly affected by feedback from $\mathcal{P}$. The adversarial objective of the discriminator is:

$$\mathcal{L}_{\mathcal{D}} = \mathbb{E}_{\theta\sim\mathbb{P}_\theta}[(\mathcal{D}(\theta) - 1)^2] + \mathbb{E}_{z\sim\mathbb{P}_z}[\mathcal{D}(\mathcal{G}(z))^2] \tag{10}$$

where $\mathbb{P}_\theta$ is the real pose distribution of the AMASS dataset.

Animatable NeRF

We use animatable human NeRF (Su et al. 2021) (A-NeRF) to render human images in new poses and from new viewpoints. A-Nerf enables rendering a human body in unseen poses and unseen viewpoints. The merit of A-NeRF compared with classical rendering methods is that it does not require human 3D scans and enables rendering personalized human images (with A-NeRF trained on personalized data).

Given a sequence of frames from a person $[\mathbf{I}_k]_{k=1}^N$, A-NeRF is aimed to optimize 3D poses $[\theta_k]_{k=1}^N$ and a parameterized body model C_ϕ. ϕ and θ are optimized for an image reconstruction objective as follows:

$$\mathcal{L} = \Sigma||C_\phi(\theta_k) - I_k||_1 + \lambda_\theta d(\theta_k - \hat{\theta}_k) + \lambda_t||\frac{\partial^2\theta}{\partial t}||. \tag{11}$$

The last term applies a smoothness prior and the middle term enforces the optimized θ to be close to $\hat{\theta}$ estimated by a 3D pose estimator. We render synthetic human images via ray marching as follows:

$$C(u,v;\theta_k) = \Sigma_{i=1}^{Q} T_i(1 - exp(-\sigma_i\delta i))c_i, \tag{12}$$

where (u,v) are 2d location in the image and T is defined as $T_i = exp(-\sum_{j=1}^{i-1}\sigma_j\delta_i)$. σ_j is the volume density at sample location j along the ray and δ_i is the distance between two adjacent points along the ray.

Pose Estimator

The pose estimator is trained with the image and 3D pose pairs $(\hat{I}, \theta)$. The $\hat{I}$ is rendered by A-NeRF and θ is generated by $\mathcal{G}$. We assume that θ would be the ground truth 3D poses of the rendered image. However, the A-NeRF model is not perfect and there might be some errors in the rendered image specifically for complicated poses from a novel camera viewpoint. Therefore, we add a simple constraint on the loss of $\mathcal{P}$ to exclude samples with large errors. The $\mathcal{L}_\theta$ used for training the pose estimator is $\mathcal{L}_\theta = f(||\theta_i - \hat{\theta}_i||_2)$ where $\hat{\theta}$ is the estimated pose and f is:

$$f(w) = \begin{cases} w & \text{if } w < d \\ 0 & \text{otherwise} \end{cases}. \tag{13}$$

In the above formula, d is a threshold to exclude samples with large errors. For the generalizability of our method, we

use the same loss for fine-tuning any pose estimator. The additional losses used for prior working during their pre-training process are not used for fine-tuning.

Experiments

Datasets. We use three datasets that are not used during pre-training of pose estimator $\mathcal{P}$, including 3DPW, AGORA, and SKI-Pose, for evaluation. To further evaluate the effectiveness of the proposed *PoseGen* on boosting the performance of $\mathcal{P}$ on IND datasets used in the pre-training procedure, we also perform an evaluation on 3DHP. Below we give the details of each dataset.

- **3DPW** (von Marcard et al. 2018) includes in-the-wild images of two subjects performing different tasks including climbing, boxing, and playing basketball. We use the test set of 3DPW for evaluation.
- **AGORA** includes realistic high-quality synthetic data from more than 150 subjects with varied clothing and with complex realistic backgrounds. AGORA includes frequent occluded images that make a unique dataset for evaluation of the generalizability of pre-trained models.
- **SKI-Pose** (Spörri 2016) is captured in a ski resort from 5 professional athletes. SKI-Pose includes camera viewpoint and poses rarely seen in the training of pose estimators. We use the test set of SKI-Pose for evaluation.
- **MPI-INF-3DHP (3DHP)** includes data from 8 subjects. The test set includes data from two of the subjects and includes in-the-wild and in-the-lab data. The training set of 3DHP has been used for training and the test set is used for evaluation as IND data.

Pose Estimators. Our framework can be used to fine-tune any pre-trained pose estimator. We choose two popular pre-trained 3D human pose and mesh estimator models, namely HybrIK (Li et al. 2021) and SPIN (Kolotouros et al. 2019). SPIN is a famous pre-trained model widely used as a baseline in recent works (Liu, Kortylewski, and Yuille 2023) and HybrIK is a recent method that has specifically shown promising results on cross-dataset evaluations. HybrIK is pre-trained on 3DHP, Human3.6M (Ionescu et al. 2014), and MSCOCO (Lin et al. 2015). SPIN has been trained on Human3.6M, 3DHP, and LSP (Johnson and Everingham 2010). We fine-tuned these pre-trained models with our framework and evaluated them on unseen datasets.

NeRF Model. The NeRF model (Su et al. 2021) has been trained on 1500 synthetic 3D poses from (CMU 2020). The 3D poses were rendered by (Varol et al. 2017) from 9 different camera viewpoints. The total number of images in the training set was 10800 512×512 images. We keep the NeRF model frozen during training and data generation.

Evaluation Metrics. Following previous work, we use mean-per-joint position error (MPJPE) and mean-per-joint position error after Procrustes alignment (PA-MPJPE) with the ground truth 3D poses. We also report PCK on the 3DHP dataset.

Quantitative Results

Tables 1, 2, and 3 show the evaluation results of *PoseGen* on AGORA, 3DPW, and SKI-Pose under scenario 2 assumptions. In the upper section of the tables, we show the results of pose estimators after fine-tuning (FT) on the designated dataset. In the lower section of the tables, we show the evaluation results of the pre-trained model (PT) on the test set of the designated dataset. The difference between the results of FT and PT models indicates how much the dataset is OOD for the PT models. The difference between FT and PT on AGORA, 3DPW, and SKI-Pose are about 61 mm, 17mm, and 73mm, respectively. Therefore, the test-set of AGORA and SKI-Pose are highly OOD for PT HybrIK. We show the results of SPIN and HybrIK after fine-tuning with *PoseGen* in the last part of Tables 1-3. *PoseGen* relatively improves HybrIK for about 4% and 5% in terms of MPJPE on AGORA and 3DPW, respectively. The improvements of SPIN are 15% and 7% on AGORA and 3DPW, respectively.

Table 4 shows the results of our method on the 3DHP dataset. Since 3DHP has been used for pre-training of HybrIK and SPIN, we consider that as an IND dataset. *PoseGen* improves SPIN and HybrIK for about 3% in terms of MPJPE. Therefore, our method is also effective in improving the baseline pose estimators on IND datasets. Although the 3DHP dataset has already been used during pre-training, the test set of 3DHP is challenging and involves novel poses. Therefore, pre-trained HybrIK obtains an MPJPE of 99.3 mm on 3DHP while it obtains an MPJPE of 88.7 mm on the 3DPW dataset.

Table 3 shows the results on the SKI-Pose dataset. The definition of joints (e.g. hip and shoulders) and bone lengths of ground truth 3D poses of the SKI dataset are a little bit different from SMPL joint definitions. Moreover, most of the poses of SKI-Pose contain athletes bent toward the side and from novel viewpoints. Therefore, the errors reported on SKI-pose are greater compared with other datasets. However, *PoseGen* still improves HybrIK and SPIN on the SKI-Pose dataset. Comparing all four benchmarks, our method is effective in improving baseline models.

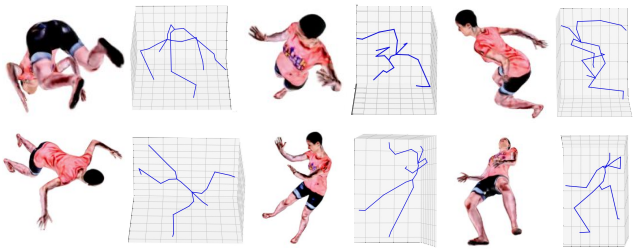

Figure 3: Sample images and 3D poses generated by *PoseGen*. Images are from novel poses and novel camera viewpoints. The 3D poses and images qualitatively are good, even though images are rendered from rare viewpoints.

Qualitative Results

Figure 3 shows some samples of generated data by *PoseGen* and their 3D poses. Images are from novel camera viewpoints that are not present in the datasets used for pre-training of the pose estimators. Most of the public datasets

Method	FT	MPJPE↓	NMJE↓
SPIN (Kolotouros et al. 2019)	✓	153.4	199.2
BEV(Sun et al. 2022)	✓	105.3	113.2
CLIFF (Li et al. 2022)	✓	81.0	89.0
HybrIK (Li et al. 2021)	✓	77.0	84.6
SPIN(Kolotouros et al. 2019)		175.1	223.1
EFT (Joo et al. 2021)		165.4	203.6
VIBE(Kocabas et al. 2020)		146.2	174.0
HybrIK(Li et al. 2021)		137.9	166.1
Ours+SPIN		**158.9**(-16.2)	**189.2**(-34)
Ours+HybrIK		**132.7**(-5.2)	**159.9**(-6.2)

Table 1: Results on AGORA dataset. Models that are trained on the training set of AGORA are shown with a checkmark.

Method	FT	PA-MPJPE↓	MPJPE↓
EFT (Joo et al. 2021)	✓	55.7	–
VIBE (Kocabas et al. 2020)	✓	51.9	82.9
HybrIK (Li et al. 2021)	✓	41.8	71.3
CLIFF (Li et al. 2022)	✓	43.0	69.3
(Li and Pun 2023)		76.8	-
SPIN(Kolotouros et al. 2019)		59.2	96.9
PoseAug (Gong et al. 2021)		58.5	94.1
VIBE (Kocabas et al. 2020)		56.5	93.5
(Choi et al. 2022)		51.5	93.5
HybrIK (Li et al. 2021)		49.3	88.7
Ours+SPIN		**56.2**(-3.0)	**89.7**(-7.2)
Ours+HybrIK		**48.3**(-1.0)	**84.4**(-4.3)

Table 2: Results on 3DPW dataset. Models that are trained on the training set of 3DPW are shown with a checkmark.

Method	FT	PA-MPJPE↓	MPJPE↓
(Rhodin et al. 2018)	✓	-	85
(Wandt et al. 2021)**	✓	89.6	128.1
SPIN(Kolotouros et al. 2019)	✓	57.2	94.3
SPIN(Kolotouros et al. 2019)		135.5	288.9
HybrIK (Li et al. 2021)		125.5	205.2
Ours+SPIN		**130.6**(-5)	**250.9**(-33)
Ours+HybrIK		**124.5**(-1)	**204.2**(-1)

Table 3: Results on SKI-Pose dataset. * Trained using multi-view cameras. ** Trained using multi-view cameras and partial 3D annotations.

Method	FT	PCK↑	MPJPE↓
HMR (Kanazawa et al. 2018)	✓	72.9	124.2
SPIN(Kolotouros et al. 2019)	✓	76.4	105.2
HybrIK (Li et al. 2021)	✓	80.0	99.3
Ours+SPIN	✓	**80.9**(+4.5)	**101.7**(-3.5)
Ours+HybrIK	✓	**85.0**(+1.0)	**96.7**(-2.6)

Table 4: Results on 3DHP dataset. All models use the 3DHP dataset for training.

	θ	K	FB	PA-MPJPE	MPJPE
Baseline				59.2	96.9
A1	✓			57.1	95.8
A2	✓	✓		56.2	92.7
A3	✓	✓	✓	55.8	91.3

Table 5: Ablation study on the main components of PoseGen.

	PA-MPJPE	MPJPE
Scenario 1	55.9	92.6
Scenario 2	56.0	91.8

Table 6: Ablation study on scenarios 1 and 2.

Distribution	PA-MPJPE	MPJPE
Uniform ($\mathcal{U}$)	55.9	91.9
Spherical ($\mathcal{S}$)	56.2	92.4
Normal ($\mathcal{N}$)	**55.8**	**91.3**

Table 7: Ablation study on prior distributions.

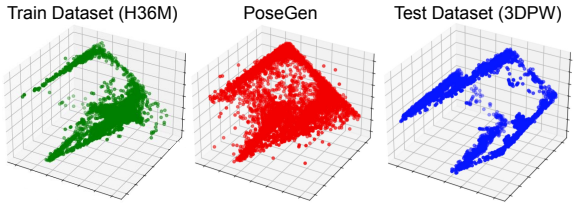

Figure 4: The distribution of camera viewpoint of the train dataset (H3.6M), the test dataset (3DPW), and the synthetic data generated by *PoseGen*.

only have chest-view cameras. Our generated data finds failure modes on top-view and bottom-view cameras and has generated such samples. Among the 4 benchmarks, the test set of AGORA is the only benchmark that covers top-view cameras. Comparing 3D poses and rendered images shows that rendered images follow the input 3D poses. Figure 4 shows the distribution of camera viewpoints in the training dataset (H3.6M), test dataset (3DPW), and synthetic data generated by our method. Our framework generates samples from unseen viewpoints, thus covering OOD viewpoints. Figure 5 shows the predictions of SPIN and PoseGen+SPIN vs. ground truth 3D poses on images from SKI-Pose and 3DPW. The top row shows that *PoseGen* improves SPIN in terms of the global orientation. Moreover, the joint angles on novel poses of athletes in the ski resort are better predicted.

Ablation Studies

Components of *PoseGen*. Table 5 shows the results of *PoseGen* after excluding the main components of the framework. In *A1* we only generate novel poses and render poses which improved the baseline for 1.1 mm (MPJPE). In *A2* we generate both novel poses and camera-viewpoint that further improves A1 for 3 mm. Adding feedback in *A3* improves *A2* for 1mm. The ablation study shows that all components are critical in improving the baseline models. Moreover, generating data from novel camera viewpoints has a major impact on improving the robustness of pre-trained models. This is well-aligned with the findings of prior work that generate 2D-3D pose pairs (Gholami et al. 2022b).

Scenarios. Table 6 compares the results of *scenario 1* and *scenario 2* on 3DPW. In *scenario 2* and *scenario 1* we obtained an MPJPE of 91.8 and 92.6, respectively. These results show that generating OOD samples in *scenario 2*

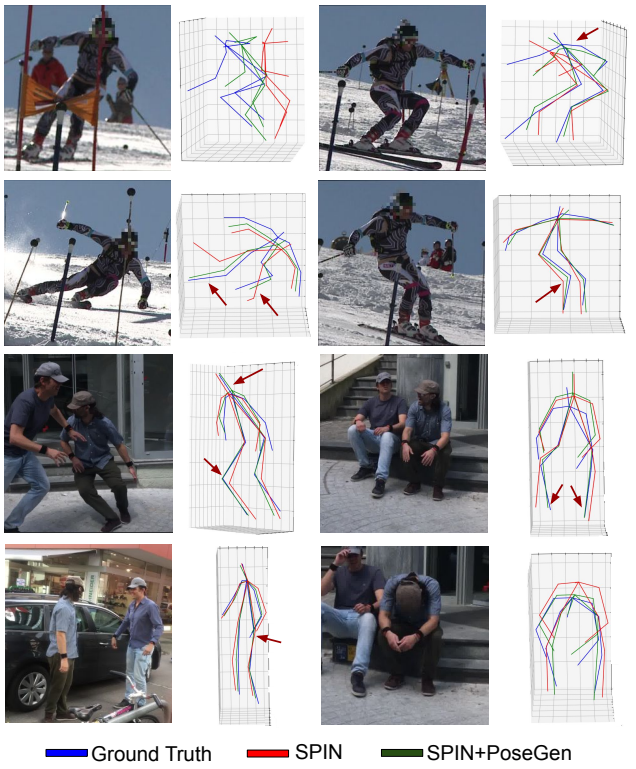

Figure 5: Predictions of SPIN and *PoseGen(+SPIN)* on in-the-wild images from SKI-Pose and 3DPW. On the challenging SKI-Pose dataset, *PoseGen* improves the predictions in the z-direction (depth) and global rotation of the human body.

makes the pre-trained model more robust on unseen OOD samples.

Prior Distributions. Table 7 shows the ablation study on different prior distributions. Previous work has shown that uniform and spherical distributions are more effective in learning a general prior for the human pose. Our experiments show that normal distribution is a better choice for our framework. Generating data with normal, uniform, and spherical distribution results in MPJPE of 91.3, 91.9, and 92.4, respectively on the 3DPW dataset. We hypothesize that our method is aimed at finding the failure modes of the pre-trained model. In contrast to prior works that try to find a smooth uniform prior, failure modes are usually discontinuous. Therefore, having a uniform distribution ($\mathcal{U}$ and $\mathcal{S}$) is not a proper prior for *PoseGen*.

Dataset Size. We increased the number of generated samples from 1K to 9k, and our experiments showed in Figure 6 that we could obtain the best performance with only 6K samples. *PoseGen* improves the performance of SPIN from 59.2 mm to 55.8 mm in terms of PA-MPJPE with only 6K samples. AGORA (Patel et al. 2021) improves the performance of SPIN on the 3DPW dataset to 55.8 mm by generating a dataset of 14K images (each including multiple subjects) with more than 350 subjects. Therefore, our method is more

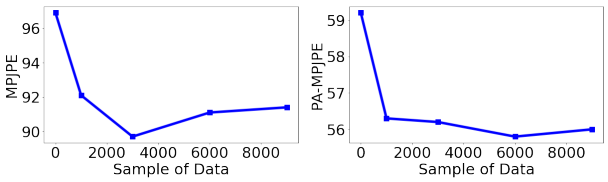

Figure 6: Performance improvement while increasing the number of generated samples.

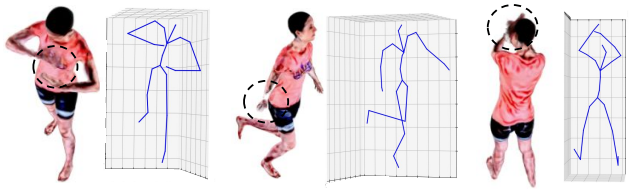

Figure 7: Failure models of *PoseGen*. A-NeRF is not perfect in rendering images from the novel viewpoint and novel poses.

efficient compared with competitors.

Limitations

The NeRF model used in this study has some limitations in rendering complex poses. Figure 7 shows some failure cases in rendering images from novel poses. The A-NeRF model is only capable of rendering images for a single subject. In order to render images for different subjects, we need to use separate NeRF checkpoints trained on that specific subject data. Future work should extend the experiments by generating a dataset that includes different subjects. We expect that having more subjects in the framework will further improve the performance.

Conclusion

In this work, we proposed an end-to-end framework for generating a 3D human pose dataset using NeRF. Our experiments showed that NeRFs are capable of generating datasets to improve the robustness of the pre-trained model. NeRFs can be trained on use-specific images and therefore the proposed framework can be used in future work to generate user-specific datasets. We performed experiments on two scenarios where we generated data 1) to minimize, and 2) to maximize the loss of a pre-trained model. We showed that the second scenario results in better performances. Moreover, we performed experiments on the prior distributions and showed that uniform and spherical prior distributions are not appropriate for the specific objective of this work.

References

Black, M. J.; Patel, P.; Tesch, J.; and Yang, J. 2023. BEDLAM: A Synthetic Dataset of Bodies Exhibiting Detailed Lifelike Animated Motion. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 8726–8737.

Choi, H.; Moon, G.; Park, J.; and Lee, K. M. 2022. Learning to Estimate Robust 3D Human Mesh from In-the-Wild Crowded Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1475–1484.

CMU. 2020. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/.

Davydov, A.; Remizova, A.; Constantin, V.; Honari, S.; Salzmann, M.; and Fua, P. 2022. Adversarial parametric pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10997–11005.

Epic Games. 2017. Unreal Engine. https://www.unrealengine.com.

Gabeur, V.; Franco, J.-S.; Martin, X.; Schmid, C.; and Rogez, G. 2019. Moulding humans: Non-parametric 3d human shape estimation from single images. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2232–2241.

Gholami, M.; Rezaei, A.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022a. Self-supervised 3D human pose estimation from video. *Neurocomputing*, 488: 97–106.

Gholami, M.; Wandt, B.; Rhodin, H.; Ward, R.; and Wang, Z. J. 2022b. AdaptPose: Cross-Dataset Adaptation for 3D Human Pose Estimation by Learnable Motion Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13075–13085.

Gong, K.; Zhang, J.; and Feng, J. 2021. PoseAug: A Differentiable Pose Augmentation Framework for 3D Human Pose Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8575–8584.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. 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.

Johnson, S.; and Everingham, M. 2010. Clustered pose and nonlinear appearance models for human pose estimation. In *bmvc*, volume 2, 5. Aberystwyth, UK.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end Recovery of Human Shape and Pose. arXiv:1712.06584.

Kolotouros, N.; Pavlakos, G.; Black, M. J.; and Daniilidis, K. 2019. Learning to Reconstruct 3D Human Pose and Shape via Model-fitting in the Loop. In *Proceedings of the IEEE International Conference on Computer Vision*.

Li, H.; and Pun, C.-M. 2023. CEE-Net: Complementary End-to-End Network for 3D Human Pose Generation and Estimation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(1): 1305–1313.

Li, J.; Xu, C.; Chen, Z.; Bian, S.; Yang, L.; and Lu, C. 2021. Hybrik: A hybrid analytical-neural inverse kinematics solution for 3d human pose and shape estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3383–3393.

Li, S.; Ke, L.; Pratama, K.; Tai, Y.-W.; Tang, C.-K.; and Cheng, K.-T. 2020. Cascaded Deep Monocular 3D Human Pose Estimation With Evolutionary Training Data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Z.; Liu, J.; Zhang, Z.; Xu, S.; and Yan, Y. 2022. Cliff: Carrying location information in full frames into human pose and shape estimation. In *European Conference on Computer Vision*, 590–606. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Bourdev, L.; Girshick, R.; Hays, J.; Perona, P.; Ramanan, D.; Zitnick, C. L.; and Dollár, P. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312.

Liu, Q.; Kortylewski, A.; and Yuille, A. L. 2023. PoseExaminer: Automated Testing of Out-of-Distribution Robustness in Human Pose and Shape Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Patter Recognition*, 672–681.

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 *2017 international conference on 3D vision (3DV)*, 506–516. IEEE.

Mehta, D.; Sotnychenko, O.; Mueller, F.; Xu, W.; Sridhar, S.; Pons-Moll, G.; and Theobalt, C. 2018. Single-Shot Multi-Person 3D Pose Estimation From Monocular RGB. In *3D Vision (3DV), 2018 Sixth International Conference on*. IEEE.

Patel, P.; Huang, C.-H. P.; Tesch, J.; Hoffmann, D. T.; Tripathi, S.; and Black, M. J. 2021. AGORA: Avatars in Geography Optimized for Regression Analysis. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Pavlakos, G.; Zhu, L.; Zhou, X.; and Daniilidis, K. 2018. Learning to estimate 3D human pose and shape from a single color image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 459–468.

Rhodin, H.; Spörri, J.; Katircioglu, I.; Constantin, V.; Meyer, F.; Müller, E.; Salzmann, M.; and Fua, P. 2018. Learning Monocular 3D Human Pose Estimation From Multi-View Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rogez, G.; and Schmid, C. 2016. MoCap-guided Data Augmentation for 3D Pose Estimation in the Wild. In Lee, D.; Sugiyama, M.; Luxburg, U.; Guyon, I.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

Rong, Y.; Liu, Z.; Li, C.; Cao, K.; and Loy, C. 2019. Delving Deep Into Hybrid Annotations for 3D Human Recovery in the Wild. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5339–5347. Los Alamitos, CA, USA: IEEE Computer Society.

Spörri, J. 2016. Research Dedicated to Sports Injury Prevention – the 'Sequence of Prevention' on the example of Alpine Ski Racing. *Habilitation with Venia Docendi in "Biomechanics". Department of Sport Science and Kinesiology, University of Salzburg.*

Su, S.-Y.; Yu, F.; Zollhöfer, M.; and Rhodin, H. 2021. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34: 12278–12291.

Sun, Y.; Liu, W.; Bao, Q.; Fu, Y.; Mei, T.; and Black, M. J. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13243–13252.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning From Synthetic Humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

von Marcard, T.; Henschel, R.; Black, M.; Rosenhahn, B.; and Pons-Moll, G. 2018. Recovering Accurate 3D Human Pose in The Wild Using IMUs and a Moving Camera. In *European Conference on Computer Vision (ECCV)*.

Wandt, B.; Rudolph, M.; Zell, P.; Rhodin, H.; and Rosenhahn, B. 2021. CanonPose: Self-Supervised Monocular 3D Human Pose Estimation in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13294–13304.

Xu, Y.; Zhu, S.; and Tung, T. 2019. DenseRaC: Joint 3D Pose and Shape Estimation by Dense Render-and-Compare. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 7759–7769. Los Alamitos, CA, USA: IEEE Computer Society.