Pose-Guided 3D Human Generation in Indoor Scene

Minseok Kim, Changwoo Kang, Jeongin Park and Kyungdon Joo*

Artificial Intelligence Graduate School, UNIST
{hello96min, kangchangwoo, jeonginpark, kyungdon}@unist.ac.kr

Abstract

In this work, we address the problem of scene-aware 3D human avatar generation based on human-scene interactions. In particular, we pay attention to the fact that physical contact between a 3D human and a scene (*i.e.*, physical human-scene interactions) requires a geometrical alignment to generate natural 3D human avatar. Motivated by this fact, we present a new 3D human generation framework that considers geometric alignment on potential contact areas between 3D human avatars and their surroundings. In addition, we introduce a compact yet effective human pose classifier that classifies the human pose and provides potential contact areas of the 3D human avatar. It allows us to adaptively use geometric alignment loss according to the classified human pose. Compared to state-of-the-art method, our method can generate physically and semantically plausible 3D humans that interact naturally with 3D scenes without additional post-processing. In our evaluations, we achieve the improvements with more plausible interactions and more variety of poses than prior research in qualitative and quantitative analysis. Project page: https://bupyeonghealer.github.io/phin/.

1 Introduction

Visual understanding of humans, from 2D human keypoint detection (Li et al. 2019a; Khirodkar et al. 2021; Jin et al. 2020), skeleton estimation (Jiang, Camgoz, and Bowden 2021) to 3D human mesh estimation (Kanazawa et al. 2018), has been actively studied in both academic and industry fields for several decades. With the advent of deep learning, visual understanding of humans has shown promising results and its applicability has been proved by interests of AR/VR companies such as Meta. In particular, generating a 3D human avatar[1] in 3D space has started to gain a lot of attention as a medium describing a human and communicating with others in the coming metaverse era. To generate a natural 3D human avatar in a given scene, it is essential to consider scene context information as well as the kinematically feasible pose of 3D human avatars. Based on parametric human model (Loper et al. 2015), recent works on 3D human generation use scene context information, such as semantic or depth maps of a scene (Zhang et al. 2020a,b). These studies also exploit physical rules to avoid collision and interpenetration with scene objects, which enables to generate plausible 3D human avatars with the scene. Inspired by this fact, we pay attention to physical contacts between 3D human avatars given a scene. Specifically, when we interact with various objects in indoor environments, contact areas of 3D human and scene objects are geometrically aligned. In addition, the contact areas of the human body in contact with the environment change depending on the posture of the 3D human. For example, when we sit on a chair, our thigh mainly contacts the chair or when we lie on a bed, the back of our body in contact with the bed. We observed that this geometrically aligned physical contact could be a clue for natural 3D human generation given a scene.

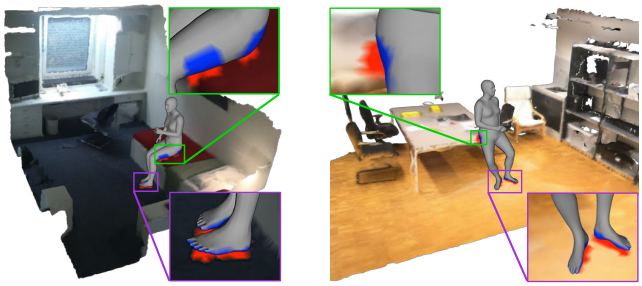


Figure 1: Examples of the proposed pose-guided 3D human generation. We visualize such geometric interactions using different colors; blue color indicates a part of the 3D human avatar that interacts with the scene, and red color denotes the opposite case. Green and purple boxes are enlarged view of scenes and human avatars to show interaction.

Based on these observation, we propose a pose-guided 3D human generation framework that considers geometrically aligned physical contacts between 3D human avatars and scene context information (see Fig. 1). Concretely, we leverage two types of geometric alignments; close distance and surface normal alignment on potential contact areas between 3D human avatars and given a scene. It encourages making physically feasible contacts of a 3D human avatar with a scene. In addition, we introduce a compact yet effective pose classifier that classifies the pose of the generated hu-

*Corresponding author.

[1]We will interchangeably use terms such as 3D human avatar, human mesh, and human body to describe 3D humans in a scene.

man avatar and provides potential contact areas of the human body parts. Thus, depending on the pose of the 3D human avatars, we can adaptively enforce the geometric alignments within the conditional Variational Autoencoder (cVAE)-based 3D human generation framework (Zhang et al. 2020b) in the form of geometric alignment loss. We show that our method can generate physically and semantically plausible 3D humans that interact naturally with 3D scenes without additional post-processing. In particular, our method results in improvements that have higher diversity metrics, gain physical plausibility metrics and more suitably interact with surroundings in quantitative experiments. In summary, our main contributions are as follows:

- We propose a pose-guided 3D human generation framework conditioned on a scene. Through the pose-guided model, generated 3D human avatars are geometrically aligned with scenes in terms of feasible contacts.
- We present a compact yet effective pose classifier to classify generated human avatar poses. According to the classified pose of the human avatar, we adaptively concentrate on different potential contact areas between 3D human avatars and 3D scenes for geometric alignment.
- We propose a geometric alignment loss, which jointly leverages the potential contact area between 3D human avatars and a given scene. It enables close distance and surface normal alignment on potential contact areas.

2 Related Work

Human Affordance Prediction. Affordance is defined as the relationship between human and object, more broadly, scene. It means that the possible set of actions that an actor can perform should concern surrounding environment (Hassanin, Khan, and Tahtali 2018, 2021). Affordance can be considered in various tasks that involve the visual understanding of the human, such as hand pose estimation (Grady et al. 2021; Corona et al. 2020; Williams and Mahapatra 2019), 3D human avatar generation (Li et al. 2019b; Zhang et al. 2020a; Hassan et al. 2021), 3D pose generation (Wang et al. 2019), motion prediction (Huang et al. 2022; Cao et al. 2020; Wang et al. 2021), object affordance prediction (Do, Nguyen, and Reid 2018; Kim and Sukhatme 2014; Fang et al. 2018), and shape estimation (Clever et al. 2020).

We aim to answer the question of how a 3D human avatar can be placed appropriately in a 3D indoor scene while recognizing its surroundings. We believe that when humans interact with scenes, surface normal and distance are related to placing a human geometrically correctly.

Scene-Aware Human Mesh Generation. The development of methods to populate 3D human avatars in 3D scenes has received considerable attention in recent years (Kim et al. 2014; Li et al. 2019b; Zhang et al. 2020b; Wang et al. 2021; Hassan et al. 2021) Putting humans into feasible locations in a scene is a complicated task. Various body poses should be considered, and humans should be placed without collision with the surroundings. Li *et al.* (Li et al. 2019b) propose a 3D pose generative model to place 3D body skeletons into the input scene represented by depth image, RGB, or RGB-D. Zhang *et al.* (Zhang et al. 2020a) use Basis Point Sets (BPS) (Prokudin, Lassner, and Romero 2019) to encode the relationship between human and objects a given 3D scene. Hassan *et al.* (Hassan et al. 2021) predict which part of the body is in contact with an object given a fixed posture. In particular, Zhang *et al.* (Zhang et al. 2020b), our baseline, bring up the idea that placing 3D people in scenes will be useful for numerous applications such as securing training data for human pose estimation, video games, and VR/AR. They use the Chamfer distance between generated human mesh and scene to properly induce interaction between them. Unfortunately, only distance-based human-scene interaction sometimes causes undesirable effects, *e.g.*, initially generated 3D avatar on the wrong side of a scene, occurring collision when there are small parts of the body such as feet or hands, or not sitting upright in a chair. To avoid those unexpected effects, they perform additional post-processing. For the purpose of alleviating this limitation, we exploit distance measurement with normal alignment as geometric alignment. In addition, we propose a pose-guided network focused on building a 3D human generation model with generalization capabilities for unseen scenes and various postures geometrically aligned.

3 Proposed Approach

In this section, we propose a new 3D human generation framework that utilizes pose-guided human-scene interaction in a geometric manner. The overall architecture of the proposed approach is illustrated in Fig. 2. Before describing details, we briefly recap the human and scene representations in Sec. 3.1. We then present how we exploit pose-guided human-scene interaction for generating geometrically plausible 3D human avatars in Sec. 3.2. In Sec. 3.3, we explain the proposed architecture and loss function.

3.1 Representation

Human Representation We utilize the SMPL-X model (Pavlakos et al. 2019) to represent the 3D human body. SMPL-X is a differentiable function that maps from a set of low-dimensional body parameters to a 3D human body mesh. The SMPL-X representation is composed of the translation $t \in \mathbb{R}^3$, which is defined by 3D vector in meters, the rotation $R \in \mathbb{R}^6$, which is defined by a 6D continuous rotation feature (Zhou et al. 2019), the body shape parameter $\beta \in \mathbb{R}^{10}$, the body pose parameter $\theta_b \in \mathbb{R}^{32}$, which is defined in the latent space of VPoser (Pavlakos et al. 2019), which is a VAE trained on a large motion capture dataset, AMASS (Mahmood et al. 2019), and the hand pose parameter $\theta_h \in \mathbb{R}^{24}$, which is parameterized the poses of the left and right hands, respectively. We denote this human representation as $x_h := (t, R, \beta, \theta_b, \theta_h)^\top \in \mathbb{R}^{75}$.

In addition, the SMPL-X model has a fixed body topology, which consists of $10,475$ vertices and $20,908$ faces. We additionally annotate the local body parts to geometrically align humans with 3D scenes. Concretely, unlike (Zhang et al. 2020b) that divide the body into 8 body parts, we divide the fixed body topology into 26 body parts. For example, we split the thigh into a left thigh and a right thigh to independently use geometric alignments depending on the

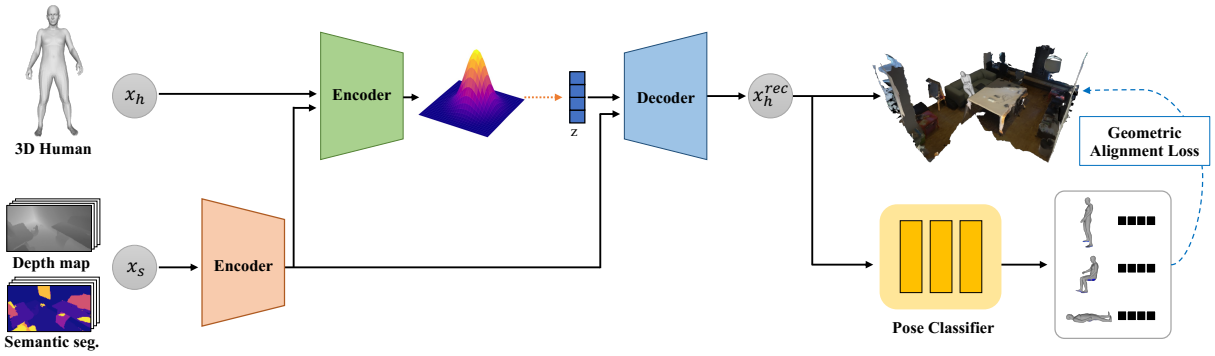


Figure 2: Overview of the proposed framework. Our framework generates a 3D human avatar in a given 3D scene using pose-guided human-scene interaction. During training, we encode a stack of depth maps and semantic segmentations x_s and encode the human avatar x_h into the latent space. The latent variable z is sampled with the VAE re-parameterization trick (Kingma and Welling 2014) (orange dashed arrow). Given the latent variable z and scene context information, we generate the 3D human avatar x_h^{rec}. In addition, the pose classifier takes x_h^{rec} as an input and then classifies its pose and provides potential contact areas of the generated 3D human avatar, which allows us to apply our geometric alignment loss adaptively.

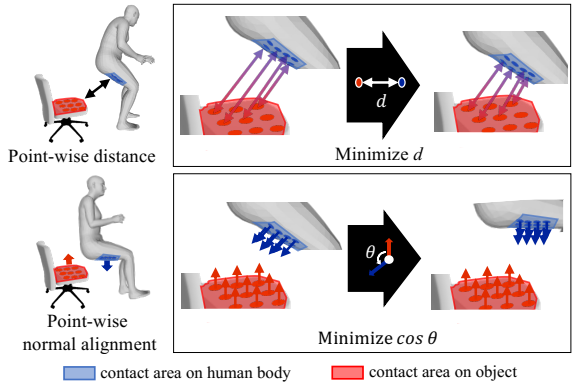


Figure 3: Illustration of how geometric alignment loss works. *Left*: Overall appearance of the human avatar and the object for each geometric alignment loss (*i.e.*, distance and normal losses). *Right*: Illustration of the before/after effect of minimizing each geometric loss. Distance loss computes the point-wise distance between the human avatar and the object of the scene and operates in the direction of minimizing it. Normal loss operates in the direction of minimizing the cosine similarity between the normal vector of human avatars and the normal vector of objects.

body parts. We use these body parts for pose-guided human-scene interaction in Sec. 3.2.

Scene Representation To encode scene context information, we use the semantic and depth data stack as the scene representation as in (Zhang et al. 2020b). We denote the scene representation as x_s, the camera perspective projection from 3D to 2D as $\pi(\cdot)$, and an inverse perspective projection from 2D to 3D as $\pi^{-1}(\cdot)$. We normalize the 3D coordinates to the range of $[-1, 1]$ using $\pi(\cdot)$. Furthermore, we additionally extract surface normal vectors of 3D scenes for geometric alignment. To place a human avatar in a 3D scene, the camera extrinsic parameter, T_c^w, transforms the 3D human body mesh coordinates to the world coordinates.

3.2 Pose-Guided Human-Scene Interaction

Scene-aware human generation aims to generate plausible poses of human avatars given a scene (Zhang et al. 2020b). In this task, valid human-scene interaction (*i.e.*, avoiding collisions or interpenetration) is critical for generating realistic 3D human avatars, where we can explain this physically proper interaction as feasible contact between 3D human and scene (see Fig. 1).

To this end, we exploit geometric alignment on contact between human and scene in terms of physically valid human-scene interaction. In addition, we adaptively consider this geometric alignment according to human pose, that is, pose-guided human-scene interaction. Thus, we propose a compact yet effective pose classifier for guidance. It allows us to generate more realistic human avatars.

Geometric Alignment on Potential Contact Area Inspired by 3D point cloud registration (Rusinkiewicz and Levoy 2001), we utilize two types of geometric alignments on potential contact areas: point-wise distance and point-wise surface normal alignment. Given potential contact areas between the human body and the scene object, point-wise distance allows us to make the human body close to the scene object. Point-wise surface normal alignment enables correct alignment, as shown in Fig. 3.

Specifically, as point-wise distance, we minimize the sum of the Chamfer distance:

$$\mathcal{L}_{dist} = \frac{1}{|P_b|}\left(\sum_{\mathbf{p}_i^b \in P_b} \min_{\mathbf{p}_i^s \in P_s} \|\mathbf{p}_i^b - \mathbf{p}_i^s\|_2 + \sum_{\mathbf{p}_i^s \in P_s} \min_{\mathbf{p}_j^b \in P_b} \|\mathbf{p}_i^s - \mathbf{p}_j^b\|_2\right), \tag{1}$$

where P_b is the set of vertices on the selected human body part, P_s is the set of point clouds on the scene, and $|\cdot|$ denotes the cardinality. For each vertex point $\mathbf{p}_i^b \in P_b$, the Chamfer

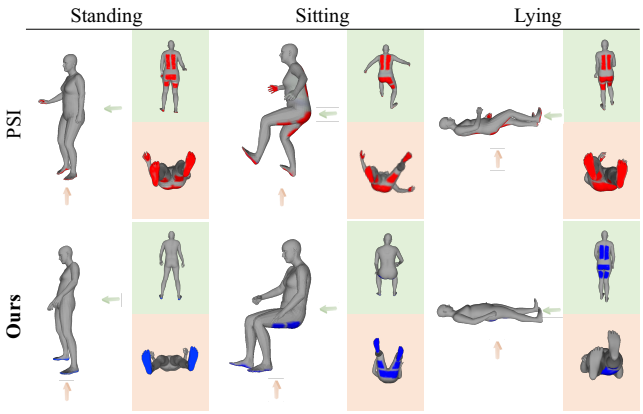


Figure 4: Illustration of contact areas of human body parts. We illustrate the potential contact areas of body parts according to the posture of 3D human avatars. *Top*: Regardless of posture, prior research (Zhang et al. 2020b) uses fixed body parts (red color regions). *Bottom*: Depending on the posture of 3D human avatars, contact areas of human body parts vary in ours (blue color regions), which enables adaptive geometric alignment.

distance finds the nearest point $\mathbf{p}_i^s \in P_s$ by measuring the Euclidean distance, and as for each scene point $\mathbf{p}_i^s \in P_s$ vice versa. Note that for the selected body part P_b, we limit the corresponding scene points P_s to the nearest scene points with respect to P_b. Without this constraint, let the nearest scene point of a vertex on feet be a point p_k^s on a chair seat, the nearest body vertex of p_k^s can be on thighs, for example.

To align point-wise surface normal, we minimize the cosine distance between the surface normal vector of the selected body vertex and the surface normal vector of the scene point. If a body part and scene object are in contact, the surface normals of the body part and scene object should be parallel ideally. Thus, the cosine similarity between them is π (or zero). Our surface normal alignment loss is:

$$\mathcal{L}_{normal} = \sum_{i=1}^{|P_b|} \left(1 + \frac{\langle \mathbf{n}_i^b, \mathbf{n}_i^s \rangle}{\|\mathbf{n}_i^b\|_2 \|\mathbf{n}_i^s\|_2}\right), \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product, and $\mathbf{n}_i^b$ and $\mathbf{n}_i^s$ are the surface normal vectors of $\mathbf{p}_i^b$ and $\mathbf{p}_i^s$, respectively. We preprocess the surface normal outward from the central mass.

It should be worth noticing that if we consider only point-wise distance as (Zhang et al. 2020b) did, it may cause interpenetration between generated human body and scene object, as shown in Fig. 6.

Pose Classification for Guidance Although considering geometric alignment between the human avatar and scene helps to generate a plausible pose of the human avatar, one thing to keep in mind is that humans have numerous body poses. We observed that the contact areas of the human body change according to human pose. For example, when we are sitting, our feet and thighs contact scene objects, such as chairs and sofas, but our thighs do not generally reach scene objects when standing (see Fig. 1). Based on this observation, we propose a compact yet effective pose classifier that classifies a given pose of humans and provides guidance in applying geometric alignment.

We design the pose classifier as an MLP-based network. Our pose classifier takes as input SMPL-X parameters x_h representing the pose of the generated human and then classifies it into four categories: standing, sitting, lying, and ambiguous poses. According to the classified poses, we can provide potential contact areas for geometric alignment in an adaptive manner (see Fig. 4). Note that the contact area of the ambiguous pose is a combination of contact areas of standing, sitting, and lying. We perform a separate pose classification task using our PROX-P dataset, which is an additionally annotated dataset based on the PROX dataset (Hassan et al. 2019); more details are provided in Sec. 4.1.

3.3 Architecture and Loss Function

Network Architecture The proposed network is built upon a cVAE-based human generation approach (Zhang et al. 2020b) that generates the human avatars conditioned on a given scene. Overall, during the training, the latent distribution learns how to generate a geometrically plausible human given training scene using the geometric alignment loss adaptively. During the test, 3D human avatar is generated using a learned latent vector, semantic and depth scene.

Specifically, our architecture consists of encoder, decoder, and pose classifier parts, as shown in Fig. 2. Following (Zhang et al. 2020b), we have two encoder parts: human encoder and scene encoder. The scene encoder takes as input the stack of semantic and depth maps x_s (*i.e.*, scene information), and then encodes this scene information, which allows us to enforce scene-aware conditions on the human encoder and decoder. In the case of the human encoder, it takes as input SMPL-X parameters of human avatar x_h and the scene information and encodes it into a latent vector $\mathbf{z}$. Given this latent vector and scene information, the decoder generates 3D human avatar. In addition, the pose classifier, composed of an MLP-based network, classifies the SMPL-X parameters of the generated human, which provides potential contact areas to calculate the geometric alignment loss for enhancing pose-guided human generation.

Loss Function The loss function for training the proposed framework consists of two parts. One is for adaptive geometric alignment according to pose classification, and the other is for the generating part.

Pose Classification Loss $\mathcal{L}_{pose}$. For human pose classification, we use cross-entropy defined as:

$$\mathcal{L}_{pose} = -\sum_{i=1}^{N} t_i \log\left(c_i\right), \tag{3}$$

where N is the number of human pose, t_i is the ground-truth label and c_i is the softmax probability for the i-th class. We pre-train the pose classification model using $\mathcal{L}_{pose}$ and utilize the trained classifier in our generative model.

Geometric Alignment Loss $\mathcal{L}_{geo}$. The geometric alignment term encourages geometrically accurate alignment of poten-

tial contact areas between human and scene:

$$\mathcal{L}_{geo} = \alpha_{dist}\mathcal{L}_{dist} + \alpha_{normal}\mathcal{L}_{normal}, \tag{4}$$

where $\mathcal{L}_{dist}$ and $\mathcal{L}_{normal}$ denote the Chamfer distance loss in Eq. (1) and the surface normal loss in Eq. (2), respectively, and α_{dist} and α_{normal} indicate the weight factors for each loss term. Note that we adaptively use this geometric alignment loss according to the pose of the generated human avatar. Compared to utilizing whole body parts, utilizing only specific body parts based on the posture can prevent interpenetration with scenes and generate geometrically plausible 3D human avatars.

Generation Loss $\mathcal{L}_{gen}$. Following the previous work (Zhang et al. 2020b), we use the following loss term for 3D human generations:

$$\mathcal{L}_{gen} = \alpha_{kl}\mathcal{L}_{KL} + \alpha_{vp}\mathcal{L}_{VPoser} + \alpha_{coll}\mathcal{L}_{coll} + \alpha_{rec}\mathcal{L}_{rec}, \tag{5}$$

where α_{kl}, α_{vp}, α_{coll}, and α_{rec} denote the weight factors for the KL-divergence, VPoser, collision, and reconstruction losses, respectively. KL-divergence loss $\mathcal{L}_{KL}$ is given by:

$$\mathcal{L}_{KL} = D_{KL}\left(q\left(z \mid x_h\right) \| \mathcal{N}(\mathbf{0}, \mathbf{I})\right), \tag{6}$$

where $q\left(z \mid x_h\right)$ denotes the VAE encoder and $\mathcal{N}(\mathbf{0}, \mathbf{I})$ indicates the Multivariate Gaussian distribution. VPoser loss $\mathcal{L}_{VPoser}$ (Pavlakos et al. 2019) encodes natural poses with a normal distribution in latent space:

$$\mathcal{L}_{VPoser} = \left|\theta_b^{rec}\right|^2, \tag{7}$$

where θ_b^{rec} denotes the body feature of the generated human avatar. VPoser loss encourages the generated human avatar to have natural poses. Collision loss $\mathcal{L}_{coll}$ is designed to prevent conflicts with human avatars and scenes:

$$\mathcal{L}_{coll} = \mathbb{E}\left[\left|\Psi_s^-\left(T_c^w \mathcal{M}\left(x_h^{rec}\right)\right)\right|\right], \tag{8}$$

where $\mathcal{M}\left(\cdot\right)$ denotes the body mesh. We generate the body mesh from the SMPL parameter and transform it to world coordinates by T_c^w. Then, we compute the negative signed distance field (SDF) $\Psi_s^-\left(\cdot\right)$. Collision loss minimizes the mean absolute value of the negative SDF. Reconstruction loss $\mathcal{L}_{rec}$ is:

$$\mathcal{L}_{rec} = \frac{\left|x_h - x_h^{rec}\right| + \left|\pi\left(x_h\right) - \pi\left(x_h^{rec}\right)\right|}{2}, \tag{9}$$

where x_h^{rec} denotes the human representation of the generated human avatar. $\pi(\cdot)$ denotes the projected and normalized translation.

The entire training loss can be formulated as

$$\mathcal{L} = \mathcal{L}_{geo} + \mathcal{L}_{gen}, \tag{10}$$

4 Experiments

We evaluate the proposed 3D human generation framework in various aspects. Specifically, we first describe our implementation details including datasets in Sec. 4.1. We then quantitatively and qualitatively evaluate the proposed method in Sec. 4.2 and perform ablation study in Sec. 4.3.

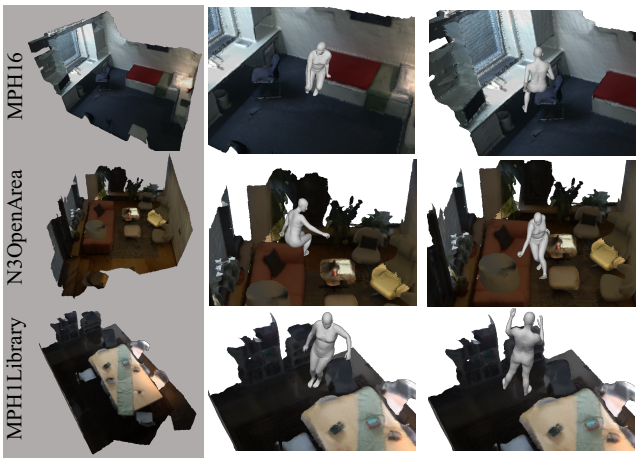


Figure 5: Generated human bodies in different test scenes of the PROX dataset (Hassan et al. 2019). Each row represents 3D human avatars generated in a given scene.

4.1 Implementation Details

Dataset We use the most widely used benchmarks in 3D human generation methods for evaluation: the PROX dataset (Hassan et al. 2019) and the PROX-E dataset (Zhang et al. 2020b). The PROX dataset contains various and natural actions of 3D human model, represented by the SMPL-X model (Pavlakos et al. 2019), in different 3D indoor scenes. In addition, the PROX data contains 12 in-the-wild 3D scenes. In the case of the PROX-E dataset, which is an extended version of the PROX dataset, it additionally provides scene information such as semantic and depth maps in image domain, and downsampled scene point clouds and Signed Distance Function (SDF) in 3D domain. Following (Zhang et al. 2020b), we use 'MPH16', 'MPH1Library', 'N0SittingBooth' and 'N3OpenArea' as test scenes, and the rest of scenes for training. For more details on both datasets, we refer to (Zhang et al. 2020b).

For the 3D human pose classification, we modify the PROX and PROX-E datasets in terms of human pose, and we call it PROX-Pose, in short the *PROX-P* dataset. Concretely, we add additional three types of annotations: the pose of 3D human in PROX videos, the body part segments, and the surface normal vector of scenes. We extract frames from human capture 8 videos in PROX scenes every 0.1 seconds and then manually annotate the human pose into 4 classes. In detail, we set up the additional class for ambiguous pose except for sitting, standing, and lying. We obtain about 12.5k pose annotations in total. A 10k data is utilized for training, while a 2.5k data is used for testing. For the annotation of the local body parts, we mainly considered body areas that come into contact with the scene. We divide human point clouds into 26 body parts carefully, unlike (Zhang et al. 2020b). For example, we separately annotated the human thigh considering left/right and front/back. Our body segments are refined and better suited for learning kinematic relationships in 3D space. As for the surface normal of scene point clouds, we extract the surface normal vectors of 3D scenes. Addition-

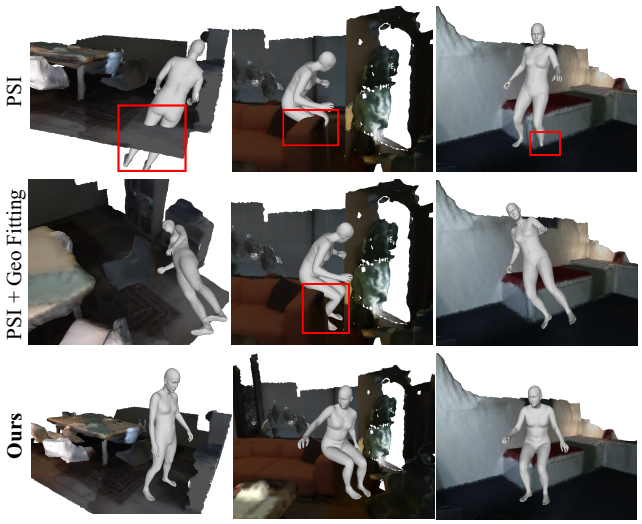


Figure 6: Qualitative comparison with PSI (Zhang et al. 2020b). The first and second rows show the generated 3D human by PSI without and with scene geometry-aware fitting, respectively. The last row shows the result of our method. Red rectangles denote penetration between human and scene.

ally, we refine the extracted surface normal by using neighbor points (Tombari, Salti, and Stefano 2010) to deal with holes in the 3D scene. Our data will be available for research.

Implementation To train our model, we use about 70k frames in the training scenes, including semantic and depth maps, scene point clouds and SDF, and the ground truth SMPL-X parameters of humans (same as in (Zhang et al. 2020b)). Concretely, we use 128×128 size of semantic and depth images to encode scene data. Our framework uses 68k point clouds of uniformly downsampled scene data and the corresponding SDF for computing the geometric alignment loss. We adopt three fully-connected layers to classify human poses and pre-train this classification model in advance.

We set $\{\alpha_{dist}, \alpha_{normal}\} = \{0.002, 0.003\}$ for the geometric alignment loss. For the generation loss, we set $\{\alpha_{kl}, \alpha_{vp}, \alpha_{coll}, \alpha_{rec}\} = \{1, 0.001, 0.1, 0.001\}$, where α_{kl} increases linearly in an annealing scheme (Bowman et al. 2015) for training. We use the Adam optimizer (Kingma and Ba 2014) with the learning rate $3e^{-4}$. The batch size is set to 32. Our model is trained 30 epochs, which takes around 1 day. All experiments are implemented in Pytorch v1.7.1 (Paszke et al. 2019) with Nvidia RTX 3090 GPU.

4.2 Evaluation

Comparison Methods We compare our method with three state-of-the-art methods in 3D human generations. PSI (Zhang et al. 2020b), which is our baseline, generates 3D human given the scene depth and the semantic segmentation. PLACE (Zhang et al. 2020a) encodes scene-human relationship using a basis point set (BPS) representation (Prokudin, Lassner, and Romero 2019), which is a sparse distance information. POSA (Hassan et al. 2021) uses human-centric formulation given a fixed human pose.

Method	non-coll	contact	entropy	cluster size
PSI	0.94	0.99	2.97	2.53
PLACE	0.98	0.99	2.91	2.72
POSA	0.97	**1.0**	2.94	2.28
Ours	**1.0**	**1.0**	**2.98**	**4.69**

Table 1: Evaluation of physical plausibility and diversity metrics.

Quantitative Evaluation We adopt the same quantitative evaluation metrics used in (Zhang et al. 2020b,a; Hassan et al. 2021). In addition, we evaluate the pose classifier using the pose prediction accuracy.

Physical Plausibility. We evaluate the non-collision and contact scores between the generated body mesh and scene mesh, defined by Zhang *et al.* (Zhang et al. 2020b). The non-collision score is the ratio of body mesh vertices with positive SDF values divided by the total number of SMPL-X vertices. The contact score is 1 if at least one vertex of the human body mesh has a non-positive scene SDF value. We report the average non-collision score and the average contact score of 4,800 samples. PSI, PLACE, and POSA are comparable under these metrics. Table 1 shows the results on the physical plausibility metric, where the arrows next to the metric indicate the direction of better performance. Thanks to pose-guided alignment, our method achieves a higher non-collision score and contact score compared to the comparison methods, which indicates that our proposed method can geometrically align between human avatar and scenes. Note that the PROX-P data is only used for training the pose classifier, not for generation.

Diversity Metric. This metric evaluates how diverse the generated human bodies are. We compute the diversity metric using $4,800$ generated SMPL-X sampling data. Like (Zhang et al. 2020b), we perform K-means clustering ($K = 20$) to cluster the SMPL-X parameters of the generated bodies. We evaluate the entropy of the cluster ID histogram of all the samples and the average size of all the clusters. The higher is the better for both metrics. Table 1 shows the diversity metric. Overall, our method outperforms the baseline. Remarkably, our method significantly enhances the average cluster size, which indicates the generated human avatars have a wide range of poses and positions in the scene. We believe that this is because our method considers the kinematic relationship between the human pose and the scene.

Pose Classifier. We evaluate the pose classifier on the PROX-P dataset including annotated human poses. In the 2.5k test set, we achieve 98% classification accuracy. We deduce that since we properly simplify the pose of 3D human avatars into four feasible categories, we can get high classification performance with an even compact classification network. In addition, by virtue of this high classification accuracy, we can adaptively and robustly exploit geometric

Method	non-coll	contact	entropy	cluster size
Distance	**1.0**	**1.0**	2.94	2.65
Normal	**1.0**	**1.0**	2.94	2.63
Ours	**1.0**	**1.0**	**2.98**	**4.69**

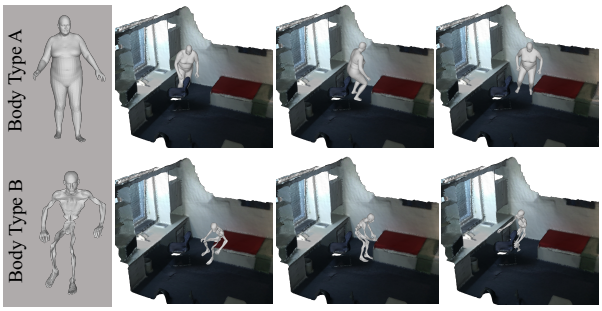
Table 2: Ablation study on physical plausibility and diversity metrics.

alignments depending on the pose of the 3D human avatar.

Qualitative Evaluation Figure 5 shows generated 3D human avatars in different test scenes. We can observe various postures of 3D human avatars, geometrically aligned with the scene in test scenes. Our model achieves competitive visual performance. Note that our results are directly generated by our method without any independent optimization-based refinement, unlike PSI using post-processing for refinements. However, despite the proper geometric alignment between the scene and the human avatars, our method may generate the human avatars penetrating the object.

In addition, we perform a more detailed comparison with PSI. In Fig. 6, the first and second rows show the generated 3D human avatars by PSI without and with post-processing, respectively. The last row shows the result of our method. When PSI generates plausible 3D humans, the post-processing (*i.e.*, refinement step) improves the quality of the generated 3D human, as shown in the last row in Fig. 6. However, since PSI uses only distance-based interactions regardless of the pose of 3D human, it sometimes generates interpenetration. In addition, the post-processing can rather make the generated 3D human worse than before, as shown in the first two rows in Fig. 6. In contrast, our method robustly generates plausible 3D human avatars without additional post-processing, since we consider the geometric alignment between human pose and scene using the pose-guided framework.

Generalization Performance We validate the generalization performance of our method from two perspectives in Fig. 7. We first experiment with how our method works depending on different body types. We modify the SMPL-X parameters associated with body shape to generate different body shape avatars. In particular, we used height and weight to find specific parameters of the desired body type using Virtual Caliper software (Pujades et al. 2019). Although we did not use different body shapes, such as thin or fat shapes, for training, our method shows reasonable results, as shown in Fig. 7(a). We believe that geometric alignment, especially surface normal alignment, enables us to generate plausible 3D humans with even different body shapes. We also apply our method to unseen scene data to show the possibility of generalization performance. For an unseen 3D scene from the MP3D-R dataset (Chang et al. 2017), the proposed approach generates plausible 3D human avatars (see Fig. 7(b)).



(a)

(b)

Figure 7: Generalization performance. (a) Results with different body shapes, such as thin or fat shapes. (b) Results on the *unseen MP3D-R* data (Chang et al. 2017).

4.3 Ablation Study

We analyze the influence of the geometric alignment loss to validate the effectiveness of our method. In this case, we use the physical plausibility and diversity metrics. As shown in Table 2, we present model performances when trained with or without distance loss $\mathcal{L}_{dist}$ and normal loss $\mathcal{L}_{normal}$ in Eq. (10), respectively. We found that the loss term $\mathcal{L}_{dist}$ in training increase the cluster size over normal loss $\mathcal{L}_{normal}$. Notably, when combining the distance loss $\mathcal{L}_{dist}$ with the normal loss $\mathcal{L}_{normal}$, we can observe that the cluster size is significantly improved.

5 Conclusion

We propose a pose-guided framework for generating a parametric model of a 3D human in an indoor scene. We explore the problem of generating 3D human while considering geometric alignment on potential contact areas between 3D human avatars and 3D scenes. Notably, by virtue of the proposed human pose classifier, we can adaptively apply geometric alignment on potential contact areas according to the classified human pose. The experimental results demonstrate the effectiveness of the pose-guided human generation networks without additional post-processing for refinement.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2022-0-00612, Geometric and Physical Commonsense Reasoning based Behavior Intelligence for Embodied AI and No.2020-0-01336, Artificial Intelligence Graduate School Program (UNIST)), the 2021 Research Fund (1.210032.01) of UNIST (Ulsan National Institute of Science & Technology), and the Ministry of Science and ICT and NIPA through the HPC Support Project.

References

Bowman, S. R.; Vilnis, L.; Vinyals, O.; Dai, A. M.; Jozefowicz, R.; and Bengio, S. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

Cao, Z.; Gao, H.; Mangalam, K.; Cai, Q.-Z.; Vo, M.; and Malik, J. 2020. Long-Term Human Motion Prediction with Scene Context. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *European Conference on Computer Vision (ECCV)*.

Chang, A.; Dai, A.; Funkhouser, T.; Halber, M.; Nießner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *International Conference on 3D Vision (3DV)*.

Clever, H. M.; Erickson, Z.; Kapusta, A.; Turk, G.; Liu, K.; and Kemp, C. C. 2020. Bodies at Rest: 3D Human Pose and Shape Estimation From a Pressure Image Using Synthetic Data. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Corona, E.; Pumarola, A.; Alenya, G.; Moreno-Noguer, F.; and Rogez, G. 2020. GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Do, T.-T.; Nguyen, A.; and Reid, I. 2018. AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Fang, K.; Wu, T.-L.; Yang, D.; Savarese, S.; and Lim, J. J. 2018. Demo2Vec: Reasoning Object Affordances From Online Videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Grady, P.; Tang, C.; Twigg, C. D.; Vo, M.; Brahmbhatt, S.; and Kemp, C. C. 2021. ContactOpt: Optimizing Contact To Improve Grasps. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hassan, M.; Choutas, V.; Tzionas, D.; and Black, M. J. 2019. Resolving 3D human pose ambiguities with 3D scene constraints. In *IEEE International Conference on Computer Vision (ICCV)*.

Hassan, M.; Ghosh, P.; Tesch, J.; Tzionas, D.; and Black, M. J. 2021. Populating 3D Scenes by Learning Human-Scene Interaction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hassanin, M.; Khan, S.; and Tahtali, M. 2021. Visual Affordance and Function Understanding: A Survey. *ACM Computing Surveys (CSUR)*, 54(3).

Hassanin, M.; Khan, S. H.; and Tahtali, M. 2018. Visual Affordance and Function Understanding: A Survey. *CoRR*, abs/1807.06775.

Huang, B.; Pan, L.; Yang, Y.; Ju, J.; and Wang, Y. 2022. Neural MoCon: Neural Motion Control for Physically Plausible Human Motion Capture.

Jiang, T.; Camgoz, N. C.; and Bowden, R. 2021. Skeletor: Skeletal transformers for robust body-pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jin, S.; Liu, W.; Xie, E.; Wang, W.; Qian, C.; Ouyang, W.; and Luo, P. 2020. Differentiable hierarchical graph grouping for multi-person pose estimation. In *European Conference on Computer Vision (ECCV)*.

Kanazawa, A.; Black, M. J.; Jacobs, D. W.; and Malik, J. 2018. End-to-end recovery of human shape and pose. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Khirodkar, R.; Chari, V.; Agrawal, A.; and Tyagi, A. 2021. Multi-Instance Pose Networks: Rethinking Top-Down Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*.

Kim, D. I.; and Sukhatme, G. S. 2014. Semantic labeling of 3D point clouds with object affordance for robot manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*.

Kim, V. G.; Chaudhuri, S.; Guibas, L.; and Funkhouser, T. 2014. Shape2pose: Human-centric shape analysis. *ACM Transactions on Graphics (TOG)*, 33(4): 1–12.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Welling, M. 2014. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*.

Li, W.; Wang, Z.; Yin, B.; Peng, Q.; Du, Y.; Xiao, T.; Yu, G.; Lu, H.; Wei, Y.; and Sun, J. 2019a. Rethinking on multi-stage networks for human pose estimation. *arXiv preprint arXiv:1901.00148*.

Li, X.; Liu, S.; Kim, K.; Wang, X.; Yang, M.-H.; and Kautz, J. 2019b. Putting humans in a scene: Learning affordance in 3d indoor environments. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 34(6): 248:1–248:16.

Mahmood, N.; Ghorbani, N.; Troje, N. F.; Pons-Moll, G.; and Black, M. J. 2019. AMASS: Archive of motion capture as surface shapes. In *IEEE International Conference on Computer Vision (ICCV)*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive body capture: 3d hands, face, and body from a single image. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Prokudin, S.; Lassner, C.; and Romero, J. 2019. Efficient learning on point clouds with basis point sets. In *IEEE International Conference on Computer Vision (ICCV)*.

Pujades, S.; Mohler, B.; Thaler, A.; Tesch, J.; Mahmood, N.; Hesse, N.; Bülthoff, H. H.; and Black, M. J. 2019. The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements. *IEEE transactions on visualization and computer graphics*, 25(5): 1887–1897.

Rusinkiewicz, S.; and Levoy, M. 2001. Efficient variants of the ICP algorithm. In *Proceedings third international conference on 3-D digital imaging and modeling*.

Tombari, F.; Salti, S.; and Stefano, L. D. 2010. Unique signatures of histograms for local surface description. In *European Conference on Computer Vision (ECCV)*.

Wang, J.; Xu, H.; Xu, J.; Liu, S.; and Wang, X. 2021. Synthesizing Long-Term 3D Human Motion and Interaction in 3D Scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Chen, L.; Rathore, S.; Shin, D.; and Fowlkes, C. C. 2019. Geometric Pose Affordance: 3D Human Pose with Scene Constraints. *CoRR*, abs/1905.07718.

Williams, X.; and Mahapatra, N. R. 2019. Analysis of Affordance Detection Methods for Real-World Robotic Manipulation. In *International Symposium on Embedded Computing and System Design (ISED)*.

Zhang, S.; Zhang, Y.; Ma, Q.; Black, M. J.; and Tang, S. 2020a. PLACE: Proximity Learning of Articulation and Contact in 3D Environments. In *International Conference on 3D Vision (3DV)*.

Zhang, Y.; Hassan, M.; Neumann, H.; Black, M. J.; and Tang, S. 2020b. Generating 3d people in scenes without people. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.