Scene-aware Generative Network for Human Motion Synthesis

Jingbo Wang[1]   Sijie Yan[1]   Bo Dai[2*]   Dahua Lin[1,3]

[1]CUHK - SenseTime Joint Lab, The Chinese University of Hong Kong

[2]S-Lab, Nanyang Technological University

[3]Centre of Perceptual and Interactive Intelligence

{wj020, ys016, dhlin}@ie.cuhk.edu.hk, {bo.dai}@ntu.edu.sg

Abstract

We revisit human motion synthesis, a task useful in various real-world applications, in this paper. Whereas a number of methods have been developed previously for this task, they are often limited in two aspects: 1) focus on the poses while leaving the location movement behind, and 2) ignore the impact of the environment on the human motion. In this paper, we propose a new framework, with the interaction between the scene and the human motion taken into account. Considering the uncertainty of human motion, we formulate this task as a generative task, whose objective is to generate plausible human motion conditioned on both the scene and the human's initial position. This framework factorizes the distribution of human motions into a distribution of movement trajectories conditioned on scenes and that of body pose dynamics conditioned on both scenes and trajectories. We further derive a GAN-based learning approach, with discriminators to enforce the compatibility between the human motion and the contextual scene as well as the 3D-to-2D projection constraints. We assess the effectiveness of the proposed method on two challenging datasets, which cover both synthetic and real-world environments.

1. Introduction

The ability to synthesize human motions is beneficial to many real-world applications, including virtual reality, film-making, and stochastic action forecasting. Previous methods [1, 2, 6, 9, 17, 20, 31, 33] for human motion synthesis often focus only on the movements of human bodies, while the scene context is neglected. Basically, people move their bodies for interacting with the outside world and are restricted by the outside world. It is hard to execute reasonable movements without observing the surrounding environment. And thus, the problem is worth further exploring.

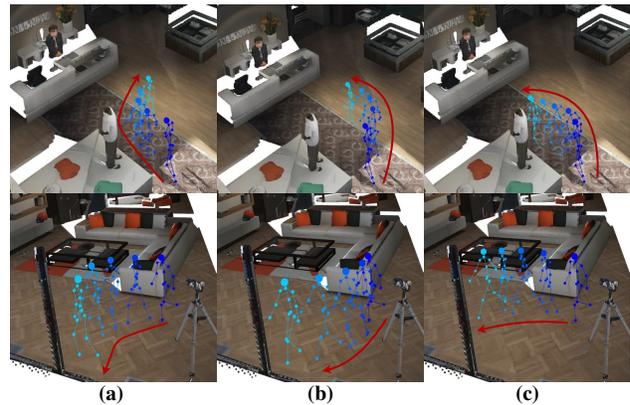Inspired by the importance of scene context, in this pa-

*Work done at The Chinese University of Hong Kong.

Figure 1. **Visualization of human motions in different scenes**. For the same starting point, the human not only can go to different goals, as in (a) and (c), but also the same goals under different trajectories and body movements, as in (a) and (b). All these human motions are sample from our generated results.

per, we aim at synthesizing human motions under scene influence. Actually, human motions in the scene consists of two components, namely body movements and the trajectory of human in the surrounding scene. This trajectory controls the human movement in the scene, and the body movements always represent the action of humans, such as walking or sitting. Thus, there are two major challenges to handle when involving scene context. The first challenge is how to effectively reflect the semantic guidance provided by the scene context, *e.g.* do sitting action on a chair. The second challenge is how to model the complicated physical relationship between scenes and action sequences. Specifically, we need to know the geometric configuration of the scene context to avoid the collision, *e.g.* where the floor is.

To solve these problems, there is an early attempt [3] introducing the scene context into motion forecasting, which supposes human actions are deterministic predictions when the history and the destination are given. This method treats the distribution of human motion as the distribution of endpoints in the scene. While in the motion synthesis task,

we argue that such this treatment may lead to gaps between the learned distribution of human motion and the one in the real world since there could be infinite ways for a person to move from one place to another. They are all valid human motions, as shown in Figure 1.

Therefore, we propose our scene-aware fully generative framework to close this gap in motion synthesis. This framework can learn the distribution of human motion in given scenes directly, rather than predicting human motions deterministically. Following [3], we represent scene context using an RGB image, which is relatively easy to acquire in real scenarios. Specifically, we divide the joint distribution as the trajectory prior in the scene and the conditional distribution of body movements given a trajectory. Inspired by the success of convolutional sequence generation networks (CSGN) [31] in skeleton-based action synthesis, we introduce the scene context into CSGN to respectively model the trajectories and fine-grained body movements. The distribution of trajectories is first learned by the trajectory generator under the condition of given scenes· Intuitively, with the guidance of the scene context and the trajectory, it is easier for the pose generator to model the distribution of semantic compatible body movements than direct synthesis (*e.g.* human always do sitting action with the static trajectory and context information of chair). In this way, our method is fully generative and is capable of capturing the diversity of human motions at various levels.

To fulfill the physical compatibility, it is crucial to introduce the geometry structure of the scene as prior knowledge into our synthesis framework. Therefore, we supervise the encoder to extract geometry context from the scene by the depth map. Under this supervision, the encoder can provide geometry-aware features of the scene, and we do not need to provide depth information of the scene during inference, which can be used more easily. Moreover, we also propose a projection discriminator and a context discriminator as geometry discriminators to further encourage the compatibility between synthesized human motions and the surrounding scene context. We deploy the projection discriminator on the 2D human motion in image coordinate space projected by the 3D human motion, because the abnormal human motion can be clearly exhibited by scale changes of humans in 2D space with the scene, as shown in Figure 4. Therefore, this discriminator encourages the generator to synthesize trajectories following the global structure of the scene, such as the floor of the scene. To prevent the collision between the synthesized human motions to the objects in the scene, we deploy this context discriminator to the relative depth sequence of the human motion to the local environment at each time step. This discriminator encourages humans to move to the correct places surrounding these objects, which are shown in Figure 5.

We choose two challenging datasets to evaluate the ef-

fectiveness of our proposed geometry-aware fully generative framework, covering both a synthesized environment (*GTA-IM* [3]) and a real environment (*PROX* [10]). On both datasets, the proposed framework is capable of synthesizing promising human motions, in terms of the fidelity of each independent sequence, the diversity of multiple sampled sequences, as well as the consistency between synthesized sequences and their corresponding scenes. To better quantitatively assess different methods, we also propose a series of new metrics for human motion synthesis with scene context, including Motion FID, which is inspired by FID for image synthesis, and Non-collision Rate, which borrows insights from 3D computer games and examines the potential collisions between human motions and the scenes.

To summarize our contributions: 1) We reformulate the task of human motion synthesis with scene context as a conditional generation problem to avoid the limitations of deterministic prediction in previous works. We further propose a series of quantitative metrics to enhance the evaluation protocol of this task. 2) We develop a novel geometry-aware fully generative framework for this task, which explicitly takes the scene geometry into consideration and captures the diversity of human motions in a scene from multiple levels. 3)We propose two geometry-aware discriminators to encourage the compatibility between synthesized human motions and their corresponding scenes.

2. Related Works

Pose Generation   Recently, rather than capturing human poses directly [25, 26, 27, 29], lots of work begin to focus on pose sequence generation. HP-GAN [1] combines the Seq2Seq model to the GAN framework for motion generation. Cai *et al.* [2] propose a Two-Stage GAN to generate the spatial and temporal information respectively for pose generation. PSGAN [33] takes the initial pose as input and action label as the condition to generate pose sequence for video generation. CSGN [31] formulates both generator and discriminator as graph convolution and generates pose sequence from noise sequence directly. Action2Motion [9] generates human pose sequences with a CVAE model for the given action. However, all these methods neglect the role of scene context in motion synthesis, and our framework is the first one for this task, as far as our knowledge.

Pose Prediction   Pose prediction is also another important task to understand human behaviors. For given continuous pose sequences, these models can predict the future human motion at a few time steps. *Encoder-Recurrent-Decoder* (ERD) [6] incorporates encoder and decoder models before and after the recurrent units for motion prediction. Based on the Seq2Seq [28] model, Martinez *et al.* [20] predicts the velocities rather than the positions of joints for motion prediction. Ac-Lstm [17] enhances the capability of LSTM
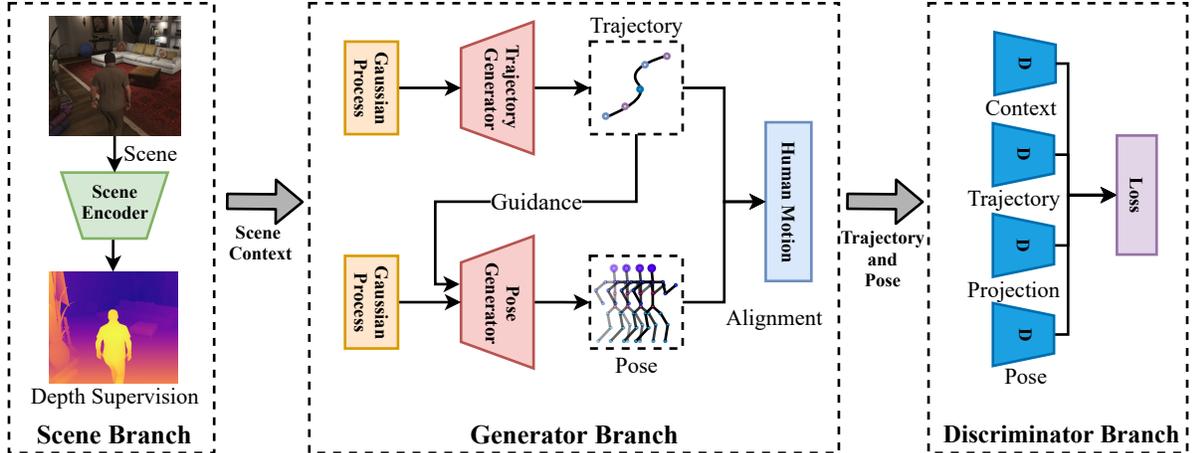
Figure 2. **Overview of our framework**. There are three components in our framework. The first component is the scene branch, which extracts the geometry-aware feature of the given scene and sends it to the generative network as the condition. The second branch is our motion generator. We first learn the distribution of trajectory, and the pose distribution is learned under the guidance of sampled trajectories. The human motion can be synthesized by aligning different sampled trajectories as well as poses. At last, the discriminator branch is introduced into this framework for the synthesized motions, and the detailed structure of this branch is demonstrated in Figure 3.

by training the mixture of synthesized frames and observed frames. Graph convolution network (GCN) is also widely used in motion prediction in recent advances [4, 15, 19]. These methods model dynamic spatial and temporal relationship from the obvious frames to the future frames. Recently, more researchers focus on human motion prediction under 2D or 3D scene context [18, 22, 23, 24, 30, 34]. Cao *et al.* [3] propose a three-stage motion prediction method which can predict different human motions under different destinations. However, our method is significantly different against previous motion prediction methods. Our fully generative network can directly learn the distribution of human in the scene and synthesize diverse human motions, as shown in Figure 1, rather than the deterministic prediction.

3. Scene-aware Generative Network

At first, we formally define the problem of human motion synthesis given the scene context. The human motion is represented as $X = (R, P)$, where R stands for the trajectory in the scene, and P stands for the pose sequence. Inspired by previous methods [3, 16, 35], we represent the scene as an RGB image. Besides the RGB image, we also provide the initial pose following previous methods [3, 9], and they together constitute the input condition S. Without loss of generality, we fix the start point of R at the image center, since we can always crop the image around the start point. Based on these notations, human motion synthesis under the given scene context thus can be described as sampling a valid X from a conditional distribution:

$$X \sim p((R, P)|S). \qquad (1)$$

Instead of learning a deterministic mapping from a given

S to some X, we propose a fully generative framework based on generative adversarial networks [7] to directly model the distribution $p((R, P)|S)$, so that the diversity of X under the given S can be fully captured. Moreover, inspired by the observation that humans usually subconsciously plan a rough trajectory in mind before moving, we divide the conditional distribution $p((R, P)|S)$ into two independent distributions as:

$$p((R, P)|S) = p(R|S) \cdot p(P|R, S), \qquad (2)$$

and further organize our framework in two stages consisting of two different GANs. These two GANs respectively capture $p(R|S)$ and $p(P|R, S)$ and are jointly trained in an end-to-end manner. Such a decomposition not only significantly reduces the computational complexity but also provides the flexibility to add trajectory- or pose-specific constraints to the proposed framework. Specifically, we include two extra discriminators, namely a projection discriminator and a context discriminator to emphasize the consideration of scene context when modeling $p(R|S)$. An overview of our final two-stage generative framework with scene-aware constraints is included in Fig. 2. Below we briefly introduce these components separately, and include their detailed architectures in the supplemental materials.

3.1. Scene Encoder

Given a scene in the form of an RGB image, a good scene encoder E_{scene} should reflect both the visual scene semantics and the scene structure in its extracted scene feature $\mathbf{f}_{scene}$, as humans are likely to move following the scene structure and perform actions that are semantically consistent with the scene, *e.g.* going upstairs when there is a stairway in the scene. To capture visual semantics of the scene,
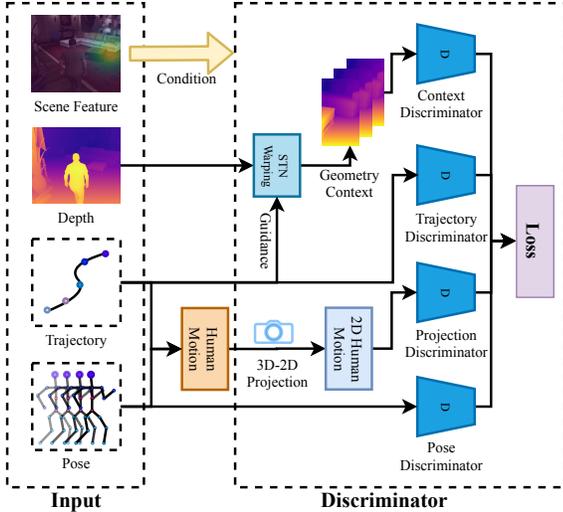
Figure 3. **Overview of our discriminator branch**. This discriminator branch contains four discriminators. The trajectory and pose discriminator help the synthesized human motions to be smooth and continuous. The projection and context discriminator encourage the human motions following the physical structure of the scene.

we deploy a ResNet-18 [11] pre-trained on ImageNet [5] as the scene encoder. More importantly, depth estimation is added as an auxiliary task for E_{scene} when training it with the whole framework jointly. To estimate the depth map from the given scene image, E_{scene} is required to include the scene structure in the extracted feature $\mathbf{f}_{scene}$. In practice, depth estimation is achieved via an inverse huber loss $\mathcal{L}_{scene}$, which is commonly adopted in recent works for monocular depth estimation [14].

3.2. Generator Branch

As mentioned, the proposed generative framework naturally decomposes $p((R, P)|S)$ into two independent distributions, namely $p(R|S)$ and $p(P|R, S)$. Subsequently, a trajectory generator is used to model $p(R|S)$ and synthesize trajectories under the condition S.

Once a trajectory R is synthesized, at each time step R provides the location in the scene, as well as the moving speed and orientation. Since all this information served as strong priors for the distribution of pose sequences, a pose generator is thus adopted to learn $p(P|R, S)$ conditioned on both S and R. Finally, we append the trajectory R to the root joint of the pose sequence P to form our final output.

3.2.1 Trajectory Generator

The trajectory generator G_{traj} represents a trajectory R of length T_R as a sequence of velocities $V = (\mathbf{v}_0, \mathbf{v}_2, ..., \mathbf{v}_{T_R-1})$ so that the location R_t at time step t is obtained via $R_t = \sum_{i=0}^{t-1} \mathbf{v}_i$. To learn V, G_{traj} follows a process where it gradually increases the sequence length to

T, based on the scene feature $\mathbf{f}_{scene}$ and the initial pose. Such a coarse-to-fine process is shown to effectively reduce the learning complexity. Specifically, we at first sample a sequence of latent vectors $Z = (\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_{T^{(0)}-1})$ where its length $T^{(0)}$ is smaller than the final length T_R, and each $\mathbf{z}$ is a $d_{\mathbf{z}}$-dimensional vector. Since trajectories should be smooth, the sequence Z is set to follow a Gaussian Process, where c-th component of $[\mathbf{z}_0, ..., \mathbf{z}_{T^{(0)}-1}]$ satisfies $[\mathbf{z}_0^{(c)}, ..., \mathbf{z}_{T^{(0)}-1}^{(c)}] \sim GP(\mathbf{0}, \kappa)$ with $\kappa(t_1, t_2) = \exp(-\frac{|t_1-t_2|}{2\sigma_c^2})$. We change the value of σ_c for each c to encourage them capture different temporal correlations. After Z is sampled, G_{traj} will use its first block $G_{traj}^{(1)}$ consisting of deconvolution layers to upsample Z into a feature sequence $F^{(1)}$ of length $T^{(1)}$ where $T^{(1)} = 2T^{(0)}$. This upsampling step will be repeated K times so that $F^{(K)} = G_{traj}^{(K)} \circ G_{traj}^{(K-1)} \circ ... \circ G_{traj}^{(1)}(Z)$ and $T^{(K)} = T_R = 2^K T^{(0)}$. While each feature in $F^{(K)}$ is a scalar, the velocity sequence V is obtained by $V = \alpha \tanh(F^{(K)})$ conducted at each time step. The hyperparameter α is used to control the average moving speed. It's worth noting the velocity sequence is modeled in the 3D camera coordinate system rather than the 2D image coordinate system to avoid potential issues raised by the 3D-2D projection.

3.2.2 Pose Generator

Conditioned the scene feature $\mathbf{f}_{scene}$, the initial pose, and the synthesized trajectory R of length T_R, the pose generator G_{pose} will output a pose sequence P consisting of $T_P = T_R$ poses each of which contains J joints. G_{pose} follows a similar structure with the trajectory generator G_{traj}. Starting from a sequence of latent vectors Z' that follows another Gaussian Process, G_{pose} gradually upsamples Z' into the pose sequence P via blocks of graph-upsampling layers, as in [31]. It's worth noting that the pose sequence is also represented in the 3D camera coordinate system, and all poses in the synthesized pose sequence are center subtracted as their movements in the scene will be controlled by the synthesized trajectory R as $X = P + R$.

3.3. Discriminator Branch

While both the trajectory generator G_{traj} and the pose generator G_{pose} have a corresponding discriminator, namely D_{traj} and D_{pose}, to distinguish synthesized trajectories and pose sequences from real ones, relying on only the scene feature $\mathbf{f}_{scene}$ to reflect the scene structure is insufficient as shown in Fig. 4, where the synthesized trajectory and pose sequence looks natural when viewed in isolation, but lead to some inconsistency when combined with the scene. Therefore, to further enhance the compatibility between synthesized human motions and the given scene, we propose two additional discriminators, namely
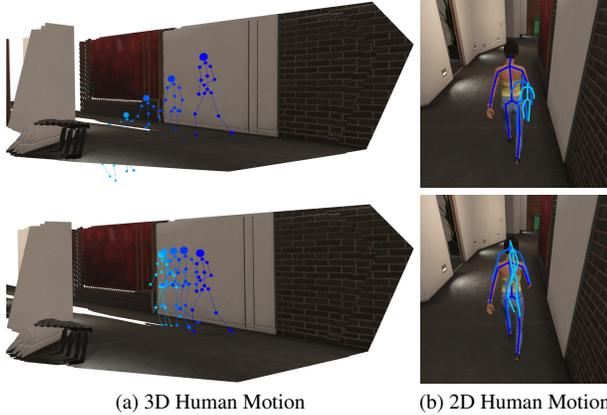
(a) 3D Human Motion	(b) 2D Human Motion

Figure 4. **Visualization of 3D human motion and the projection of this motion in 2D space**. The results in the first row is sampled from the method without projection discriminator and the second is supervised by this discriminator. With the scale changes of human motions in the 2D scene, the geometry compatibility of them can be easily judged.

the projection discriminator D_{proj} and the context discriminator D_{context}, that respectively focus on global structural constraints such as walls, the floor or the ceiling, and local structural constraints such as chairs and tables.

3.3.1 Projection Discriminator

The projection discriminator D_{proj} is used to enhance the compatibility between synthesized human motion X and the global scene structure, so that X will not result in the human goes through the wall, collides into the floor, or floats in the air. One challenge to solve when applying D_{proj} is that the motion X, either real or fake, is represented in the 3D camera coordinate system, while the scene is described using a 2D image, the mismatch between 3D and 2D thus may lead to an inferior discriminator. Fortunately, D_{proj} will only be deployed during training, and the camera matrix of each scene image is often provided by existing datasets. Therefore, D_{proj} will project the input human motion X into the image coordinate system, acquiring its 2D counterpart X'. As shown in Figure 4, the projected human motion X' can be effectively represented by the changes of 2D coordinates and the scale of human in the scene, which are relatively easier for the discriminator to make judgments. Since X' is a trajectory-aligned 2D pose sequence, D_{proj} is built upon graph convolution layers with graph downsampling, with a final global average pooling layer to aggregate information across all time steps.

3.3.2 Context Discriminator

Besides the projection discriminator D_{proj}, we also include a context discriminator D_{context} to encourage the compatibility between the synthesized human motion X and the
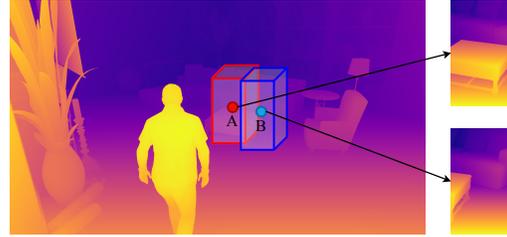

Figure 5. **Visualization of the local context for context discriminator**. Point A is the collision position to the desk and the point B is the compatible position of the scene for human motions. The geometry context of these two positions are significantly different.

local scene structure, which mainly refers to the constraint that a human should avoid hitting objects in the scene when moving. Such a constraint requires D_{context} to have a detailed understanding of the local scene context around X at each time step. On the other hand, as shown in Figure 5, when a human motion X is incompatible with the local scene structure, the relative depth between it and the scene contains meaningful patterns. Inspired by this observation, D_{context} utilizes local relative depth crops as its main source of judgments. Specifically, D_{context} will also project the input human motion X into its 2D counterpart X', followed by cropping the scene depth map around X' at each time step. The crop at t-th time step has the shape $\left(\frac{H_c}{d_t}, \frac{W_c}{d_t}\right)$, where d_t is the corresponding depth value of X' at that time, and (H_c, W_c) is the pre-defined crop size (in practice, we set it to be $\frac{1}{4}$ of the image size.). All these local depth crops will subtract their corresponding depth values of X' and resize to (H_c, W_c), forming the local relative depth crop sequence. It's worth noting the scene depth map is estimated from the scene feature $\mathbf{f}_{\mathrm{scene}}$ as discussed in Section 3.1, and the Spatial Transformer Network (STN) [12] is used as a differentiable cropping function.

4. Experiments

4.1. Dataset

In this paper, we mainly evaluate the proposed method on two public datasets, which contain both 3D human motions and geometry information of different scenes.

GTA-IM: GTA-IM [3] is a recent dataset based on the virtual GTA environment. This dataset contains long-term and diverse human motions in different indoor and outdoor scenes, as well as sufficient information of the scene, including depth maps and instance labels. We select 70 sequences in 6 scenes as the training set and 30 sequences in 3 scenes as the testing set. Then we sampled sub-sequences by a sliding window with a fixed length of 65 frames and a step of 5 frames, where the begin frame is as the condition for motion synthesis. Two scenes in the testing set are also shared with the training set, and one scene is exclusive. Although these two sets contain the sequences in the same

scene, they are synthesized in different camera viewpoints and the initial pose for generation is also different. Therefore, testing our method on same scenes is non-trivial on this dataset. Similar to [3], we only choose 19 key joints from 98 human joints provided by this dataset for human motion synthesis.

PROX: Proximal Relationships with Object eXclusion (PROX) [10] is captured by the Kinect. This dataset contains 12 different 3D scenes and provides 3D human model sequences. To evaluate our method in the real-scene, we split 10 scenes as the training set and 2 scenes as the testing set. To improve the quality of synthesized motions, we first train our model on GTA-IM and the fine-tune it on this dataset. We select 16 joints from the SMPLX [21] model and all these joints are contained in GTA-IM.

4.2. Implementation Details

We adopt Adam [13] algorithm under (0.5, 0.999) betas and 0.0002 learning rate as optimizer for all the experiments on both GTA-IM and PROX datasets. Following [31], our model is supervised by the WGAN-GP [8] framework. The gradient penalty is 10 and the gradient regression target is 0.1. We train our model 150 epochs and generate 64 frame pose sequences on both GTA-IM and PROX dataset.

4.3. Metric

Since the previous works did not propose metrics to measure the quality of scene-aware action generation, we introduce different metrics to evaluate the different characteristics of generated pose sequences, such as motion, pose, and compatibility with the given scene.

Motion FID: In CSGN [31], the well-known metric *Frméchet Inception Distance* (FID) has been extended to pose generation task. Because explicit action labels are not available in both GTA-IM and PROX, we change the encoder model from ST-GCN [32] model, which is used to compute FID score in [31], to **ERD** [6], which is a well-known LSTM based method for motion prediction. Therefore, this FID score is extend to measure the continuity and rationality of generated pose sequences by the action agnostic features. To well measure the quality of the synthesized motion in different lengths, firstly, we conduct different ERD models, who are trained based on the human motions with 16, 32, and 64 frames. Then we clip the synthesized T frames human motion to multiple short human motions. The clip lengths of these sequences are 2, 4, 8, 16, 32, and 64. The average value of the fid score is calculated based on the composition of these ERD models and the clip length as our Motion FID. Besides, we define the mean fid for 2 and 4 length as short-term value, 8 and 16 as middle-term value, and 32 and 64 as long-term value.

Non-collision Ratio: Besides the motion FID, we also propose the non-collision ratio as metric for physical com-
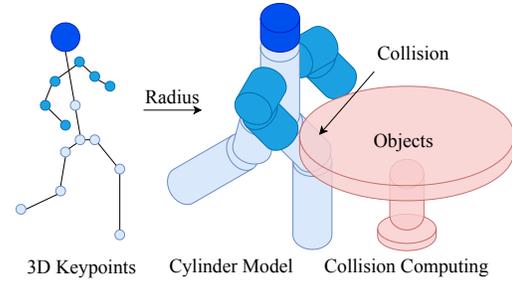

Figure 6. **Collision between human cylinder model and objects**. Pre-defined cylinder radius is conducted to the bones of the synthesized 3D human motions. We define the collision by computing the point number of 3D objects in this cylinder model.

patibility between human motions and the given scene. Firstly, we conduct cylinder model to synthesized skeleton sequences by pre-defined radius, as shown in 6. Then, the human-scene collision area can be represented as the intersection points between these cylindricalization pose sequences and the point cloud of given scenes, which are projected from the 2D depth map by camera parameters. Under this representation, the collision motion can be defined as motions with intersection points more than the pre-defined threshold t. Thus, our non-collision ratio can be calculated by the ratio between the number of human motions without human-scene collision and all sampled motions. In this work, we discuss non-collision ratio with different thresholds t: 40, 60, 80, and 100 and the radius of cylinder model with different r: 30, 45, and 60mm. We randomly sample 10000 different frames in GTA-IM and 2000 frames in PROX for ablation studies.

User Study: At last, we conduct user study to further evaluate the quality of the synthesized human motions, especially the rationality of interaction between generated human motions and the scene. In this user study, we let the users focus on the two kinds of compatibilities: one between the human motions and the given scene, and one between the trajectory and pose sequence. We named these two scores as scene compatibility and motion compatibility, respectively. For this user study, we mixed the human motions from the ground-truth, the origin CSGN, our method without geometry discriminators, and the proposed method in the same scene, and sample equal numbers of human motions for each approach. The score for this user study is between 1 (strongly abnormal) and 5(strongly compatible). And we will illustrate the mean and standard deviation of the scores from users.

4.4. Ablation Studies

In this section, we will first conduct ablation studies on GTA-IM dataset to demonstrate the effectiveness of our framework in detail. As shown in Table 1, we demonstrate the Motion FID and Non-Collision Score of the mod-
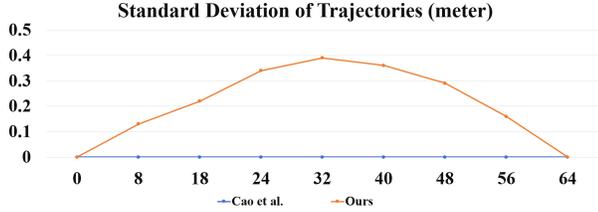
Figure 7. Standard deviation over time of 5 synthesized trajectories given the same start and end points for images in GTA-IM, where [3] obtains 0 metre since it's deterministic in this case.

Table 1. Ablation studies on GTA-IM dataset. M: separately generate trajectory and pose; D: depth supervision; P: projection discriminator; C: context discriminator.

Method	M	D	P	C	Short	Mid	Long	Ave ↓
					Motion FID			
CSGN					92.4	125.6	137.8	118.6
	✓				27.6	33.4	38.6	33.1
	✓	✓			26.8	32.2	37.5	32.1
Ours	✓	✓		✓	25.9	31.0	35.6	30.8
	✓	✓	✓		23.3	28.4	34.5	28.7
	✓	✓	✓	✓	22.4	27.6	33.7	27.6
					Non-Collision Score			
Method	M	D	P	C	30mm	45mm	60mm	Ave ↑
	✓				0.931	0.882	0.851	0.891
	✓	✓			0.941	0.902	0.874	0.902
Ours	✓	✓		✓	0.948	0.919	0.904	0.923
	✓	✓	✓		0.956	0.934	0.915	0.938
	✓	✓	✓	✓	0.962	0.943	0.924	0.943

Table 2. Comparison of Motion FID on **GTA-IM**. M: separately generate trajectory and pose; G: geometry-aware discriminator.

Method	Short	Mid	Long	Ave ↓
Two-Stage [2]	172.8	196.5	210.4	193.2
HP-GAN [1]	156.6	176.2	190.4	174.4
CSGN [31]	92.4	125.6	137.8	118.6
CSGN+ M	26.8	32.2	37.5	32.1
CSGN+ M + G	22.4	27.6	33.7	**27.6**

Table 3. Comparison of Motion FID on on **PROX**. M: separately generate trajectory and pose; G: geometry-aware discriminator.

Method	Short	Mid	Long	Ave ↓
Two-Stage [2]	261.5	276.8	292.4	276.9
HP-GAN [1]	214.7	225.3	251.7	230.6
CSGN [31]	141.5	160.3	183.2	161.6
CSGN+ M	55.3	58.4	62.5	58.7
CSGN+ M + G	52.4	54.5	56.9	**54.6**

Table 4. Collision score of the method with and without the supervision of geometry discriminator on PROX. P: projection discriminator; C: context discriminator

P	C	30mm	45mm	60mm	Ave ↑
		0.793	0.731	0.692	0.742
✓		0.813	0.764	0.733	0.770
✓	✓	0.831	0.782	0.753	**0.792**

Table 5. User studies on both GTA-IM and PROX for different methods. M: separately generative framework of human motion; G: supervision of geometry discriminators; GT: ground-truth human motions. The user study score is written as *mean(std)*.

Method	GTA-IM		PROX	
	Scene ↑	Human ↑	Scene ↑	Human ↑
CSGN	2.66(0.74)	2.92(0.84)	2.46(0.82)	2.52(0.73)
CSGN + M	3.26(0.76)	3.32(0.73)	3.12(0.88)	3.23(0.71)
CSGN + M + G	**3.56**(0.72)	**3.53**(0.68)	**3.36**(0.74)	**3.32**(0.66)
GT	4.51 (0.49)	4.45(0.49)	4.13(0.62)	4.14(0.61)

els with different module combinations. Firstly, with separation synthesis between trajectory and pose, Motion FID decreases significantly (118.6 → 33.1), which verifies the effectiveness of our design choices on generator branch. Then, the Motion FID is mitigated with discriminators for physical compatibility, and the Non-Collision score is improved from 0.902 to 0.943. The influence of depth supervision is also demonstrated in this table. The Motion FID is decreased from 33.1 to 32.1, as well as the Non-Collision score is improved from 0.891 to 0.902. Moreover, as the ablation studies on GTA-IM, we demonstrate the Motion FID and Non-collision score on PROX in Table 3 and Table 4, respectively. Our geometry-aware discriminators improve this score from 0.742 to 0.792, and the ability of them to synthesize physical compatibility motions is verified in captured real scenes. Besides, the Motion FID is mitigated from 161.1 to 54.6 based on our framework.

To further evaluate the effectiveness of proposed method, especially the design choices of our generator and discriminator, we compare our framework against different state-of-the-art methods in Table 2 and Table 3. All the methods in these tables are condition on the same scene context. The significant mitigation on Motion FID demonstrates the ability of our framework on synthesizing human motions in the scene. Besides these motion synthesis framework, we compare our framework against [3] in Figure 7. Under the same start and end points, this motion standard deviation demonstrate the diversity of sampled human motions rather than deterministic prediction.

At last, we also demonstrate the results of the user study on GTA-IM and PROX in Table 5. It is not surprising that the ground-truth human motions achieve the highest score on both datasets. For theses datasets, we observe that the geometry discriminators can significantly improves the performance for learning compatible human motions with the scene. And the two-stage human motion synthesis framework can also help the generator to synthesize smooth and continuous human motions.

4.5. Qualitative Results

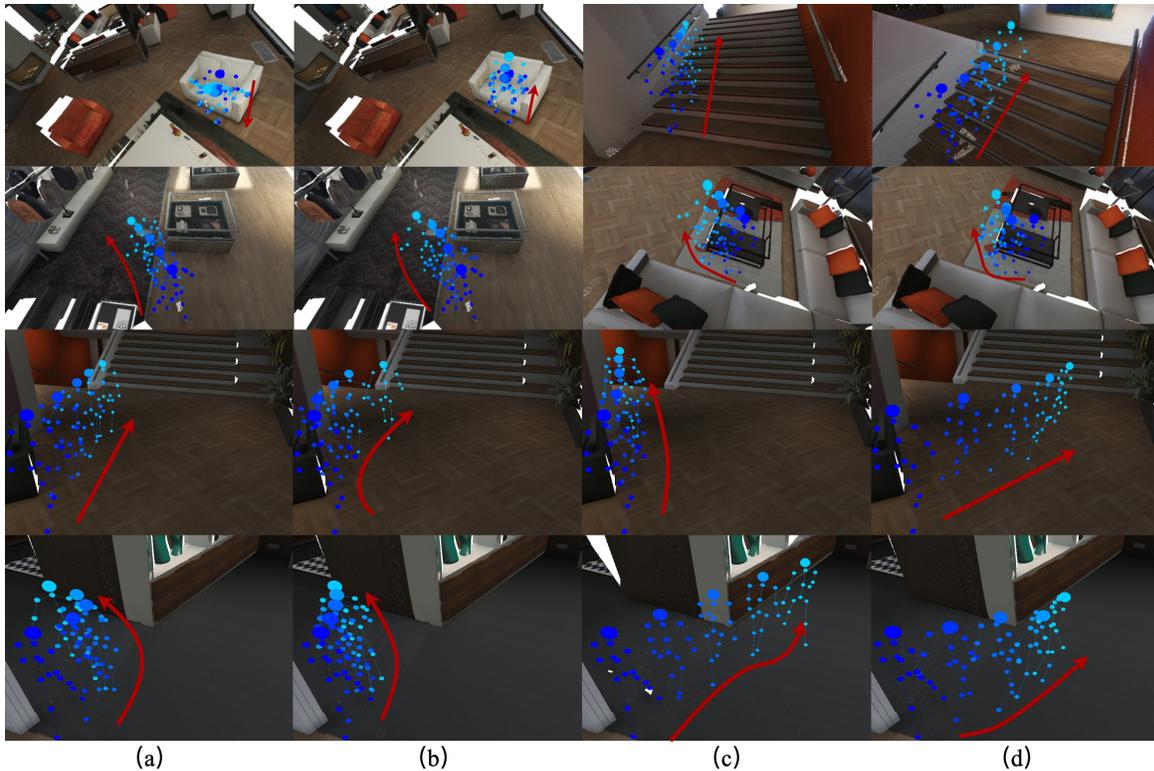At last, we show more qualitative results on the GTA-IM dataset and the PROX dataset in Figure 8 and Figure 9.

Figure 8. **Visualization on GTA-IM**. All the results are sampled from the synthesized 64 frame sequences. Our framework can synthesize diverse and compatible human motions in different scenes.
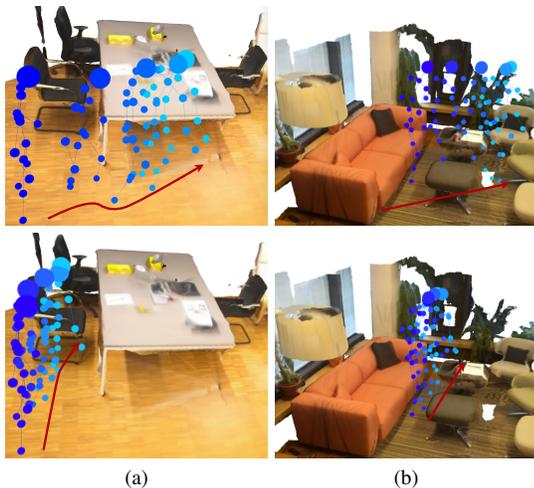

Figure 9. **Visualization on PROX**. Our method can synthesize diverse and compatible human motions in these scenes, although the environment in this dataset is more complex than GTA-IM.

From these figures, our method can synthesize diverse and compatible motions in both synthetic and real-world environments. Especially, the first row in Figure 8 indicates that our method can synthesize specific motions on chairs or stairs. And in the second row of this figure, we can also find out that the synthesized human motion turning around to avoid obstacles. The third and fourth rows demonstrate that

our method can synthesize the diverse motions with same and different end points from the same start points. More qualitative results are in the following video[1].

5. Conclusion

In this paper, we propose a novel scene-aware generative framework to model the distribution of human motion in the given scene. We at first decouple this distribution into the distribution of trajectories in the scene and the distribution of body movements based on the scene and trajectory. And a projection discriminator and a context discriminator are further introduced into our framework to encourage the compatibility between human motions and the scene. On two large-scale datasets, we demonstrate the effectiveness of the proposed generative framework which is able to synthesize human motions that are not only diverse in both trajectories and body movements, but also coherent with both structure and semantics of the given scene.

[1]Please refer to https://www.youtube.com/watch?v=XfA3QWcV0ik&t=10s for demo videos.

References

[1] Emad Barsoum, John Kender, and Zicheng Liu. Hp-gan: Probabilistic 3d human motion prediction via gan. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1418–1427, 2018. 1, 2, 7

[2] Haoye Cai, Chunyan Bai, Yu-Wing Tai, and Chi-Keung Tang. Deep video generation, prediction and completion of human action sequences. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 366–382, 2018. 1, 2, 7

[3] Zhe Cao, Hang Gao, Karttikeya Mangalam, Qi-Zhi Cai, Minh Vo, and Jitendra Malik. Long-term human motion prediction with scene context. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7

[4] Qiongjie Cui, Huaijiang Sun, and Fei Yang. Learning dynamic relationships for 3d human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[6] Katerina Fragkiadaki, Sergey Levine, Panna Felsen, and Jitendra Malik. Recurrent network models for human dynamics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4346–4354, 2015. 1, 2, 6

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 3

[8] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pages 5767–5777, 2017. 6

[9] Chuan Guo, Xinxin Zuo, Sen Wang, Shihao Zou, Qingyao Sun, Annan Deng, Minglun Gong, and Li Cheng. Action2motion: Conditioned generation of 3d human motions. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20)*, 2020. 1, 2, 3

[10] Mohamed Hassan, Vasileios Choutas, Dimitrios Tzionas, and Michael J. Black. Resolving 3D human pose ambiguities with 3D scene constraints. In *International Conference on Computer Vision*, pages 2282–2292, Oct. 2019. 2, 6

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[12] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015. 5

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2014. 6

[14] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016. 4

[15] Maosen Li, Siheng Chen, Yangheng Zhao, Ya Zhang, Yanfeng Wang, and Qi Tian. Dynamic multiscale graph neural networks for 3d skeleton based human motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[16] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12368–12376, 2019. 3

[17] Zimo Li, Yi Zhou, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. Auto-conditioned recurrent networks for extended complex human motion synthesis. *arXiv preprint arXiv:1707.05363*, 2017. 1, 2

[18] Xu Liu, Chengtao Li, Jian Wang, Jingbo Wang, Boxin Shi, and Xiaodong He. Group contextual encoding for 3d point clouds. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[19] Wei Mao, Miaomiao Liu, Mathieu Salzmann, and Hongdong Li. Learning trajectory dependencies for human motion prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9489–9497, 2019. 3

[20] Julieta Martinez, Michael J Black, and Javier Romero. On human motion prediction using recurrent neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2891–2900, 2017. 1, 2

[21] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[22] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 3

[23] Lu Qi, Shu Liu, Jianping Shi, and Jiaya Jia. Sequential context encoding for duplicate removal. *arXiv preprint arXiv:1810.08770*, 2018. 3

[24] Lu Qi, Xiangyu Zhang, Yingcong Chen, Yukang Chen, Jian Sun, and Jiaya Jia. Pointins: Point-based instance segmentation. *arXiv preprint arXiv:2003.06148*, 2020. 3

[25] Yu Rong, Ziwei Liu, Cheng Li, Kaidi Cao, and Chen Change Loy. Delving deep into hybrid annotations for 3d human recovery in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5340–5348, 2019. 2

[26] Yu Rong, Ziwei Liu, and Chen Change Loy. Chasing the tail in monocular 3d human reconstruction with prototype memory. *arXiv preprint arXiv:2012.14739*, 2020. 2

[27] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: Fast monocular 3d hand and body motion capture by regression and integration. *arXiv preprint arXiv:2008.08324*, 2020. 2

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 2

[29] Jingbo Wang, Sijie Yan, Yuanjun Xiong, and Dahua Lin. Motion guided 3d pose estimation from videos. In *European Conference on Computer Vision*, pages 764–780. Springer, 2020. 2

[30] Yajie Xing, Jingbo Wang, and Gang Zeng. Malleable 2.5 d convolution: Learning receptive fields along the depth-axis for rgb-d scene parsing. *arXiv preprint arXiv:2007.09365*, 2020. 3

[31] Sijie Yan, Zhizhong Li, Yuanjun Xiong, Huahan Yan, and Dahua Lin. Convolutional sequence generation for skeleton-based action synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4394–4402, 2019. 1, 2, 4, 6, 7

[32] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *AAAI*, 2018. 6

[33] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–216, 2018. 1, 2

[34] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018. 3

[35] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J. Black, and Siyu Tang. Generating 3d people in scenes without people. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6194–6204, June 2020. 3