2025 IEEE/SICE International Symposium on System Integration (SII)
January 21-24, 2025, Munich, Germany

Gait Sequence Upsampling using Diffusion Models for Single LiDAR Sensors

Jeongho Ahn[1], Kazuto Nakashima[2], Koki Yoshino[1], Yumi Iwashita[3] and Ryo Kurazume[2]

***Abstract*— Recently, 3D LiDAR has emerged as a promising technology in the field of gait-based person identification, serving as an alternative to traditional RGB cameras due to its robustness under varying lighting conditions and its ability to capture 3D geometric information. However, long capture distances or the use of low-cost LiDAR sensors often result in sparse human point clouds, leading to a significant decline in identification performance. To address these challenges, we propose a sparse-to-dense upsampling model for pedestrian point clouds in gait recognition using 3D LiDAR, named LidarGSU, which is designed to enhance the generalization capability of existing identification models. Our method utilizes diffusion probabilistic models (DPMs), which have shown high fidelity in generative tasks such as image completion. In this work, we leverage DPMs on sparse sequential pedestrian point clouds as conditional masks in a video-to-video translation approach, applied in an inpainting manner. We conducted extensive experiments on the *SUSTeck1K* dataset to evaluate the generative quality and recognition performance of the proposed method. Furthermore, we demonstrate the applicability of our upsampling model using a real-world dataset, captured with a low-resolution sensor across varying measurement distances.**

I. INTRODUCTION

Gait recognition is pivotal in the field of person identification. Unlike other biometric modalities, such as facial recognition, retinal scans, or fingerprints, gait offers distinct advantages, including the ability to identify individuals from a distance without requiring their cooperation, and it stands out for being non-intrusive. These unique physical characteristics make gait recognition particularly well-suited for applications in security systems and criminal investigations.

In recent years, light detection and ranging (LiDAR) has emerged as an alternative technique in the field of gait recognition, playing a critical role in mobile robotics and self-driving cars owing to its ability to capture 3D point clouds of surrounding obstacles and geometries by emitting laser beams. Previous studies on LiDAR-based gait recognition [17] have reported that it outperforms traditional RGB cameras, owing to its robustness against varying lighting conditions and capacity to provide accurate geometric information, making it more suitable for security applications. However, when deploying LiDAR sensors in identification systems, the density of human point clouds is highly sensitive to measurement distance and hardware specifications, especially when using low-resolution sensors. This sensitivity often results in significant degradation in identification performance due to the sparse or incomplete data of human shapes.

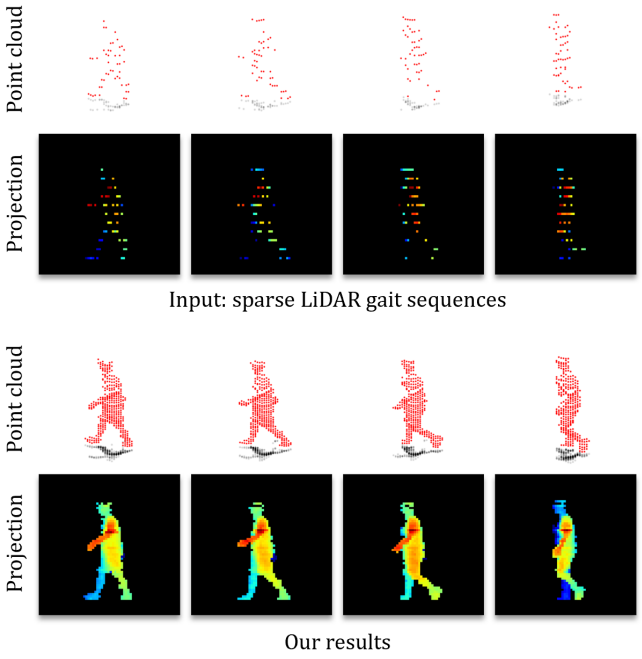

Fig. 1. Upsampled results using our models. We present sparse LiDAR gait sequence data as inputs (top two rows) alongside the corresponding outputs (bottom two rows), represented in both 3D point cloud sequences (rows 1 and 3) and 2D depth videos (rows 2 and 4).

In this paper, we present a LiDAR-based gait sequence upsampling model for sparse pedestrian point cloud data, named LidarGSU (as shown in Fig. 1), with the aim of enhancing the generalization capability of existing identification models. Our method utilizes diffusion probabilistic models (DPMs) [9], which have shown high fidelity in generative tasks, including image completion. Specifically, we treat the missing points within gait shapes using a distance-independent inpainting strategy by projecting pedestrian point clouds into a 3D Euclidean space, feeding them into a diffusion-based architecture with corresponding conditional masks. Furthermore, to ensure consistency in time-sequential gait appearances, we employed a video-based

This work was partially supported by the Japan Science and Technology Agency (JST) [Moonshot R&D][Grant Number JPMJMS2032], the JST SPRING Grant Number JPMJSP2136, and the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Number JP20H00230.

Jeongho Ahn and Koki Yoshino are with Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan {ahn, yoshino}@irvs.ait.kyushu-u.ac.jp

Kazuto Nakashima and Ryo Kurazume are with Faculty of Information Science and Electrical Engineering, Kyushu University, Japan {k_nakashima, kurazume}@ait.kyushu-u.ac.jp

Yumi Iwashita is with Jet Propulsion Laboratory, California Institute of Technology, USA yumi.iwashita@jpl.nasa.gov

noise prediction model [10] during the denoising process. To the best of our knowledge, this is the first study to address LiDAR data upsampling for gait recognition. In our experiments, we demonstrated the effectiveness of our model on two datasets: the *SUSTeck1K* dataset [17] and Ahn *et al.*'s dataset [2], both of which evaluate both generative quality and improvements in identification performance.

The contributions of this study can be summarized as follows:

- We present a LiDAR upsampling method based on conditional diffusion models that utilizes a distance-independent inpainting approach to enhance the generalization capability of existing identification models.
- By employing a video-based noise prediction technique, our diffusion model ensures consistency in the sequential pedestrian gait shapes. In addition, we used a continuous time schedule for fast and efficient LiDAR upsampling.
- In our experiments, we demonstrated that our upsampling model significantly reduces the performance gap in gait recognition tasks across LiDAR data with varying point cloud densities.

II. Related work

A. Gait recognition using LiDAR

Traditional camera-based gait recognition methods can be broadly classified into two types: appearance-based approaches and model-based approaches. The former focuses on extracting gait features directly from the visual appearances of the human body, such as images or videos [4], [5]. In contrast, the latter parameterizes visual data into human structures, such as shape-aware and non-shape-aware poses [21], [23], and analyzes them to extract gait-related features.

Most existing studies on person identification using LiDAR sensors have employed appearance-based approaches with 2D representations, as the resolution of LiDAR sensors is generally lower than that of RGB cameras, making human pose estimation less effective. Benedek *et al.* [3] proposed a projection-based model using gait energy images (GEIs) [7] to re-identify individuals in short-term scenarios. However, this method struggles to satisfactorily extract dynamic features from gait frames. Yamada *et al.* [22] utilized temporal gait changes by employing long short-term memory (LSTM) networks with sequential range representations, optimized for processing efficiency based on the sensor's specifications. However, this approach is unsuitable for real-world scenarios, such as varying capture distances and pedestrian walking directions. Ahn *et al.* [1], [2] explored view- and resolution-robust recognition frameworks by rearranging pedestrian point clouds based on a gait direction vector. Although this method enhances identification performance under complex confounding conditions, it still lacks the geometric features necessary for distance-independent analysis. Shen *et al.* [17] collected a large-scale LiDAR-based gait dataset, *SUSTeck1K*, and designed a flexible and effective projection-based identifier. While this method shows promising results compared to camera-based methods [5] at short measurement distances or with dense LiDAR sensors, such as Velodyne VLS-128, its performance degrades with longer distances or lower sensor resolutions, a challenge also noted by [22].

B. Diffusion probabilistic models for LiDAR data generation

Diffusion-based generative models [9], [18], [20] have gained significant attention across a wide range of applications, including text-to-image generation, speech synthesis, translation, and compression. Compared with generative adversarial networks (GANs) [6], another prominent generative framework, diffusion models allow for stable training with a simple objective function by approximating likelihood maximization. Specifically, denoising diffusion probabilistic models (DDPMs) [9], a type of diffusion-based model, have demonstrated notably high fidelity for various completion tasks [15], [16]. These models learn the general distribution of a dataset by iteratively adding noise to the input data and then denoising it from Gaussian noise during the inference phase.

Several studies on LiDAR data tasks, primarily focusing on scene completion, have employed diffusion-based frameworks. Zyrianov *et al.* [24] utilized NCSNv2 [19], a score-based generative model, to train both the range and reflectance modalities using image representations. Nakashima *et al.* [13] adopted unconditional DDPMs, incorporating inpainting and timestep-agnostic techniques [11], [12], to enhance both the fidelity and efficiency of LiDAR data synthesis for sim2real applications. Sander *et al.* [8] introduced LiDAR upsampling models based on conditional DDPMs [16], achieving faster sampling while maintaining high fidelity compared with prior works [13], [24]. In contrast to [8], which utilizes spherical projection, we employ conditional DPMs [16] for LiDAR gait data completion, regardless of the sensor's distance, using an orthogonal projection strategy. Furthermore, we design a video-to-video translation model to ensure the time-sequential consistency of the gait data.

III. Method

In this section, we describe the problem of addressing missing parts in gait shape sequences and introduce the formulation of conditional DDPMs in relation to LiDAR data representation, loss function, and denoiser used for iterative refinement.

A. Problem statement

The 3D point cloud data captured from single LiDAR sensors can generally be transformed into range images. Most existing studies [8], [13], [24] on LiDAR data generation have adopted a spherical projection function, which assigns an angular pixel to each angle: azimuth θ and elevation ϕ. This projection method provides well-aligned, one-to-one mapping and is cost-efficient for processing LiDAR data, as most LiDAR sensors used in autonomous driving are designed to spin mechanically and emit laser beams in a spherical pattern. In contrast, orthographic projection

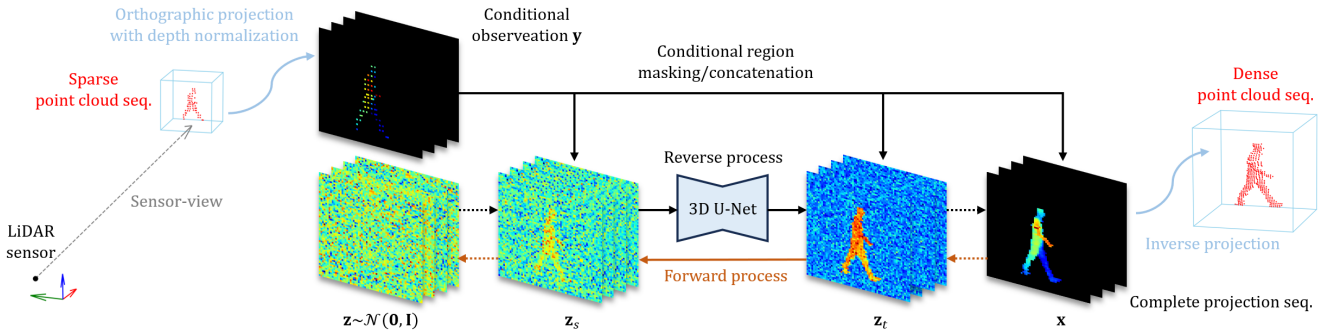

Fig. 2. Overview of our upsampling pipeline. The diffusion processes operate within the orthographic projection domain with normalized depth values. The sampled depth projection sequences can be converted back to 3D point cloud data.

directly maps LiDAR point clouds onto depth images within 3D Euclidean space. Compared to spherical projection, the orthographic projection method [1], [2], [3] preserves the full size of objects regardless of varying measurement distances and does not rely on specific laser beam patterns from the sensors. In addition, it eliminates the need for linear interpolation of pedestrian heights, which is often required in traditional camera-based gait recognition models [5], [17]. The missing points of the gait shapes in orthographic projection can be addressed using a distance- and emission-pattern-independent inpainting strategy, which is a type of linear inverse problem:

$$\mathbf{y} = \boldsymbol{H} \odot \mathbf{x}_0 + \epsilon, \tag{1}$$

where $\mathbf{x}_0$ represents a watertight gait video captured from a single LiDAR sensor, $\mathbf{y}$ is an incomplete gait video, $\boldsymbol{H}$ is a degradation noise mask, and ϵ represents noise. In this context, we assume that the noise ϵ is set to zero. Our goal is to solve this inpainting problem and recover $\mathbf{x}_0$ from the measurement $\mathbf{y}$ as a completed gait shape across varying measurement distances using conditional diffusion models.

B. LiDAR data representation

Based on the problem statement, we introduce an orthogonal projection method for LiDAR gait completion that can be directly visualized using sensors. Given a pedestrian point cloud dataset $\mathcal{P} = \{\mathcal{P}_i^j | i = 1, 2, ..., I; j = 1, 2, ..., J_i\}$ with I individuals and J_i sequences for each individual i. Each point cloud sequence $\mathcal{P}_i^j \in \mathbb{R}^{F \times N \times C}$ has F frames, N points for each frame f and the number of channels C represent Cartesian coordinates (x, y, z). Given a gait point cloud $\mathcal{P}_i^j$, we can define the center of mass $\mathbf{c}_i^j = (c_{i,f,x}^j, c_{i,f,y}^j, c_{i,f,z}^j)$ for frame f as $\mathbf{c}_i^j = \frac{1}{N}\sum_{n=1}^{N} \mathbf{p}_{i,f,n}^j$, where $c_{i,f,z}^j$ is set to zero because we only consider the sensors' emission directions on the xy-plane. Subsequently, given a sensor-view angle $\theta_{\text{sensor}}{}_{i,f}^j = \arctan(c_{i,f,y}^j, c_{i,f,x}^j)$ for the frame f on the xy-plane for a given point cloud sequence $\mathcal{P}_i^j$, we can obtain the rotated point cloud sequence $\hat{\mathcal{P}}_i^j \in \mathbb{R}^{F \times N \times C}$ with a directional angle $\theta_{\text{sensor}}{}_{i,f}^j$ as follows:

$$\hat{\mathbf{p}}_{i,f,n}^j = (\mathbf{p}_{i,f,n}^j - \mathbf{c}_{i,f}^j) \cdot \mathbf{R}_z(\theta_{\text{sensor}}{}_{i,f}^j + \pi), \tag{2}$$

where $\mathbf{R}_z$ represents the rotation matrix around the z-axis.

As in [2], we transform the point cloud sequence $\mathcal{P}_i^j$ into a gait image sequence $\mathbf{y}_i^j \in \mathbb{R}^{F \times 1 \times H \times W}$. The gait image $\mathbf{y}_{i,f}^j$ of each frame f has a resolution of $W(= l_y/r_y)$ in azimuth and $H(= l_z/r_z)$ in elevation and its depth value for an arbitrary point $\hat{\mathbf{p}}_{i,f,n}^j$ at each (h, w) is determined as follows:

$$h = \left\lfloor \frac{1}{r_z} \cdot (\hat{p}_{i,f,n,z}^j - \min_{n \in \{1,...,N\}} (\hat{p}_{i,f=0,n,z}^j) + l_{z-\text{const}}) \right\rfloor, \tag{3}$$

$$w = \left\lfloor \frac{1}{r_y} \cdot (\hat{p}_{i,f,n,y}^j + \frac{l_y}{2}) \right\rfloor, \tag{4}$$

, where l_z is the height of the z-axis, l_y is the width of the y-axis, r_z is the elevation resolution of H, r_y is the azimuth resolution of W, and $l_{z-\text{const}}$ is the z-positional normalization constant for the generated gait video $\mathbf{y}_i^j$. Here, when more than one point exists in the same pixel, the largest value is adopted, which is similar to the Z-buffer algorithm. In this work, l_z, l_y, r_z, r_y, $l_{z-\text{const}}$, H, and W are set to 2.6 m, 2.6 m, 0.04 m, 0.04 m, 0.3 m, 64, and 64, respectively.

C. Preliminaries for DDPMs

In this study, inspired by [16], we build a diffusion-based inpainting model, as shown in Fig. 2, conditioned on observation $\mathbf{y}$ (for simplicity, j and i are omitted). Additionally, we employed the DDPM framework, which formulates transitions between data and latent spaces with continuous time $t \in [0, 1]$ [11]. Compared with discrete-time diffusion models [9], a continuous noise schedule offers a finer approximation of the variational lower bound (VLB), leading to improved optimization efficiency. In standard DDPMs, the process begins with Gaussian diffusion, where the data sample $\mathbf{x}_0$ is gradually corrupted by adding Gaussian noise from $t = 0$ (least noisy) to $t = 1$ (most noisy), resulting in a noisy version of $\mathbf{x}$, referred to as latent variable $\mathbf{z}_t$.

In the forward diffusion process, the distribution of latent variable $\mathbf{z}_t$ conditioned on $\mathbf{x}_0$ for any timestep t can be given by:

$$q(\mathbf{z}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \tag{5}$$

where α_t and σ_t^2 are strictly positive scalar-valued functions of t in the noise schedule. In this study, we employed

α-cosine schedule [14], which is one of the most popular schedules, resulting in $\alpha_t = \cos(\pi t/2)$ and $\sigma_t = \sin(\pi t/2)$. Transition $\mathbf{z}_t$ can be tractably simplified using a re-parameterization trick as $\mathbf{z}_t = \alpha_t \mathbf{x}_0 + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The signal-to-noise ratio of $\mathbf{z}_t$ can be defined as $\lambda_t = \alpha_t^2/\sigma_t^2$, where $\alpha_t = \sqrt{1-\sigma_t^2}$ following the variance-preserving diffusion process [20]. The transition of the latent variable $q(\mathbf{z}_t|\mathbf{z}_s)$ from timestep s to t for any $0 \leq s \leq t \leq 1$ is also Gaussian, written as:

$$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\alpha_{t|s}\mathbf{z}_s, \sigma_{t|s}^2\mathbf{I}), \tag{6}$$

where $\alpha_{t|s} = \alpha_t/\alpha_s$ and $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2\sigma_s^2$. Given the above distributions, the reverse diffusion process $p(\mathbf{z}_s|\mathbf{z}_t)$ can be defined as:

$$p(\mathbf{z}_s|\mathbf{z}_t) = \mathcal{N}(\mu_t(\mathbf{x}_0, \mathbf{z}_t), \Sigma_t^2\mathbf{I}), \tag{7}$$

where $\mu(\mathbf{x}_0, \mathbf{z}_t) = \frac{\alpha_{t|s}\sigma_s^2}{\sigma_t^2}\mathbf{z}_t + \frac{\alpha_s\sigma_{t|s}^2}{\sigma_t^2}\mathbf{x}_0$ and $\Sigma_t^2 = \frac{\sigma_{t|s}^2\sigma_s^2}{\sigma_t^2}$.

D. Noise prediction model

In this study, we used a 3D U-Net architecture [10] as a noise prediction model $\hat{\epsilon}_\theta(\cdot)$, a parameterized neural network, to predict the noise ϵ and ensure consistent natural gait shapes across video frames F. Compared with the standard U-Net architecture in [9], [16], this 3D U-Net is factorized over space and time. Specifically, it includes space-only 3D convolution blocks, and the attention in each spatial attention block is applied over the space. In addition, temporal attention is used after each spatial attention block, with relative position embeddings applied in each temporal attention block, making it suitable for video data generation.

E. Loss function

We define the objective function for our diffusion model to estimate unknown $\hat{\mathbf{x}}$ from latent variable $\mathbf{z}_t$ with a conditional observation $\mathbf{y}$. In this paper, the target of the loss function is set to the noise ϵ, and the latent variable $\mathbf{z}_t$ for each timestep t is repeatably initialized by combining the conditional masks $\mathbf{m} \in \mathbb{R}^{F\times 1\times H\times W}$ according to observation $\mathbf{y}$ as follows:

$$\mathbf{z}_t \leftarrow \mathbf{m} \odot \mathbf{y} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{z}_t, \tag{8}$$

$$m_{(f,1,h,w)} = \begin{cases} 1, & \text{if} \quad y_{(f,1,h,w)} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Subsequently, the loss function is defined by concatenating the observation $\mathbf{y}$ and initialized latent variable $\mathbf{z}_t$ along the channel axis, as follows:

$$\mathcal{L}_{T\to\infty} = \mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(0,1)}[||\hat{\epsilon}(\text{concat}(\mathbf{y}, \mathbf{z}_t); \lambda_t) - \epsilon||_2^2]. \tag{10}$$

After the training phase, the gait data can be sampled by recursively inferring $p(\mathbf{z}_s|\mathbf{z}_t)$, where x is approximated by $\hat{\mathbf{x}}_\theta = (\mathbf{z}_t - \sigma_t\hat{\epsilon}_\theta(\text{concat}(\mathbf{y}, \mathbf{z}_t); \lambda_t))/\alpha_t$ with a finite number of timesteps T from $t = 0$ to $t = 1$. In addition, we mask the loss to compute only the unknown regions in the depth videos for more efficient training, as in [8].

IV. Experiments

In this section, we demonstrate the effectiveness of our upsampling method on both the generation and gait recognition tasks.

A. Datasets

The performance of our model was evaluated using two datasets. The first is the *SUSTeck1K* dataset [17], which is a well-known LiDAR point cloud benchmark for gait recognition. It was collected using a 128-beam LiDAR scanner (Velodyne VLS-128), capable of capturing objects and surroundings with high resolution, allowing for dense measurements. In addition, this dataset includes data from 1,050 identities, 12 gait attributes, and eight viewpoints, making it suitable for training both identification and generative models, as well as for evaluating general-purpose performance.

The second dataset [2] was collected using a 32-beam LiDAR scanner (Velodyne VLP-32C), which has a lower resolution (fewer vertical laser beams) than the sensor used in the *SUSTeck1K* dataset, resulting in sparser pedestrian point clouds at the same measurement distances, as shown in Table III. This dataset consists of 30 identities, 8 views, and 2 comparative distances, all captured with a single gait attribute (*Normal*). The rotation speed of the LiDAR sensor was the same for both datasets, operating at 10 frames per second (FPS).

B. Implementation details

Following the original protocol, we used a subset of *SUSTeck1K*, consisting of 250 subjects, for training our upsampling model. We trained our model for 200,000 iterations with a learning rate of 0.0003 and an input sequence length of 10 frames, while computing an exponential moving average (EMA) of the model weights with a decay rate of 0.995 every 10 steps. We used two types of binary noise masks to degrade the original data during the training phase: pepper noise and vertical line masks, as shown in Fig. 3. Pepper noise masks (**P**) are generated by randomly mapping points from a Bernoulli distribution to simulate noise in the azimuth based on the captured distances. In contrast, the vertical line masks (**V**) represent the beam-level noise at the elevation of the LiDAR sensors. In this study, we used three different ratios for each noise mask type and paired them with the original gait data during the training.

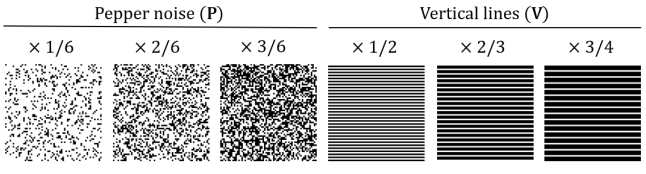

Fig. 3. Noise masks used for training and testing our model. All mask sizes are 64×64, and the black regions in each binary noise mask indicate the points removed from the clean gait data.

For the testing phase, we used the remaining test set of *SUSTeck1K*, which consists of 800 subjects, randomly

TABLE I

GENERATIVE EVALUATION OF THE SUSTECK1K DATASET WITH NOISE MASKS

Upsampling			Means (Test set)								
			$\mathbf{V}\times1/2$, $\mathbf{P}\times1/6$			$\mathbf{V}\times2/3$, $\mathbf{P}\times2/6$			$\mathbf{V}\times3/4$, $\mathbf{P}\times3/6$		
Approach	Method	Input Modality	PSNR ↑	SSIM ↑	Consistency ↓	PSNR ↑	SSIM ↑	Consistency ↓	PSNR ↑	SSIM ↑	Consistency ↓
Interpolation	Nearest-neighbor	Depth Image	6.90	0.031	0.041	6.84	0.029	0.043	6.78	0.025	0.045
Interpolation	Bilinear	Depth Image	20.90	0.852	0.016	20.99	0.841	0.017	20.83	0.840	0.019
Interpolation	Bicubic	Depth Image	21.05	0.855	0.017	21.08	0.843	0.017	20.90	0.842	0.019
Diffusion	Palette [16]	Depth Image	26.14	0.940	0.009	24.17	0.908	0.013	23.15	0.888	0.017
Diffusion	Ours w/o masking loss	Depth Video	27.22	0.953	**0.007**	25.56	**0.932**	**0.010**	24.86	**0.922**	**0.011**
Diffusion	Ours	Depth Video	**27.27**	**0.954**	**0.007**	**25.59**	**0.932**	**0.010**	**24.89**	**0.922**	**0.011**

TABLE II

IDENTIFICATION EVALUATION USING A LIDARGAIT ON SUSTECK1K DATASET WITH NOISE MASKS

Upsampling			Means (Probe set)								
			$\mathbf{V}\times1/2$, $\mathbf{P}\times1/6$			$\mathbf{V}\times2/3$, $\mathbf{P}\times2/6$			$\mathbf{V}\times3/4$, $\mathbf{P}\times3/6$		
Approach	Method	Input Modality	Rank1 ↑	Rank5 ↑	Rank10 ↑	Rank1 ↑	Rank5 ↑	Rank10 ↑	Rank1 ↑	Rank5 ↑	Rank10 ↑
			1.40	5.85	10.13	0.18	1.08	2.34	0.15	0.82	1.68
Interpolation	Nearest-neighbor	Depth Image	0.17	0.93	1.78	0.17	0.86	1.67	0.16	0.78	1.54
Interpolation	Bilinear	Depth Image	1.35	5.16	8.52	0.62	2.58	4.86	0.44	1.96	3.72
Interpolation	Bicubic	Depth Image	1.51	5.63	9.16	0.73	3.01	5.37	0.52	2.20	4.08
Diffusion	Palette [16]	Depth Image	23.62	48.69	61.07	9.93	26.61	37.31	7.16	13.79	21.82
Diffusion	Ours w/o masking loss	Depth Video	31.69	58.57	70.27	18.07	40.72	53.08	11.38	29.72	41.16
Diffusion	Ours	Depth Video	**32.49**	**59.77**	**71.28**	**18.97**	**42.09**	**54.52**	**11.85**	**30.68**	**42.26**

TABLE III

COMPARISON BETWEEN TWO DATASETS

Datasets	Sensors	Beams	V/H Res.	Subjects	Views	Average Dist.
SUSTeck1K [17]	VLS-128	128	0.11°/0.1°	1,050	12	7.5 m
Ours [2]	VLP-32C	32	1.33°/0.1°	30	8	10, 20 m

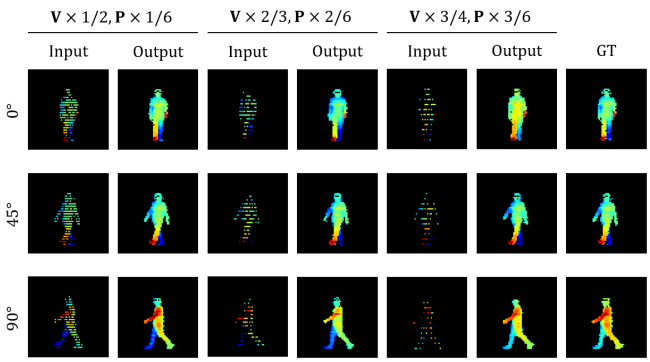

Fig. 4. Upsampled results using our model on the *SUSTeck1K* dataset for the *Normal* attribute with three noise mask combinations. The redder the color, the greater the depth value.

selecting 10 frames for each with three different combinations of noise masks. In the applicability experiment, we used Ahn *et al.*'s dataset [2]. As a baseline, we compared our diffusion model with the vanilla Palette [16] using the proposed projection on two datasets. In addition, we compared the well-established linear interpolation methods: Nearest-neighbor, Bilinear, and Bicubic. In the generative quality evaluation, we adopted the following two widely-used standard metrics: the Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). We also focused on Consistency, which is a metric for video generation that evaluates temporal coherence by calculating the gradient between consecutive video frames. In the gait recognition task, we used LidarGait [17] as the representative state-of-the-art identification model, which was pre-trained on the training set of *SUSTeck1K*, following the original protocol. In this study, identification accuracy refers to the average of the results obtained from all cross-views and gait attributes. In addition, all gallery sets consisted of clean data, whereas noise masks were applied to the probe sets during testing on the *SUSTeck1K* dataset. All diffusion-based models were configured with a fixed timestep T of 32. For this study, both our upsampling model and LidarGait, trained on the training set of *SUSTeck1K*, were applied in all experiments.

C. Generative evaluation

The quantitative generative evaluation results on the *SUSTeck1K* dataset are listed in Table I, and the examples sampled by our model for the gait attribute *Normal* are shown in Fig. 4. In Table I, diffusion-based methods significantly outperformed the interpolation approach across all three metrics. Comparing our model to Palette [16], we observed that our video-based model [10] is more effective than the image-based approach. Notably, as the noise masks became more severe, the performance gap between our model and Palette increased. For another gait attributes in Fig. 5, we can see that our method realizes high fidelity in both 2D projected gait shapes and the structure of 3D point clouds.

D. Gait recognition task

The identification results conducted on *SUSTeck1K* using LidarGait [17] are listed in Table II. In Table II, we can see that the interpolation approach achieves little to no improvement in recognition performance on sparse gait data. Similar to Table I, it can be observed that our model outperforms Palette as the noise level in the probe set increases because of its ability to maintain consistency in appearance across gait sequences, as shown in Fig. 6.

Fig. 7 shows the identification evaluation scores as functions of the number of function evaluations (NFE), which

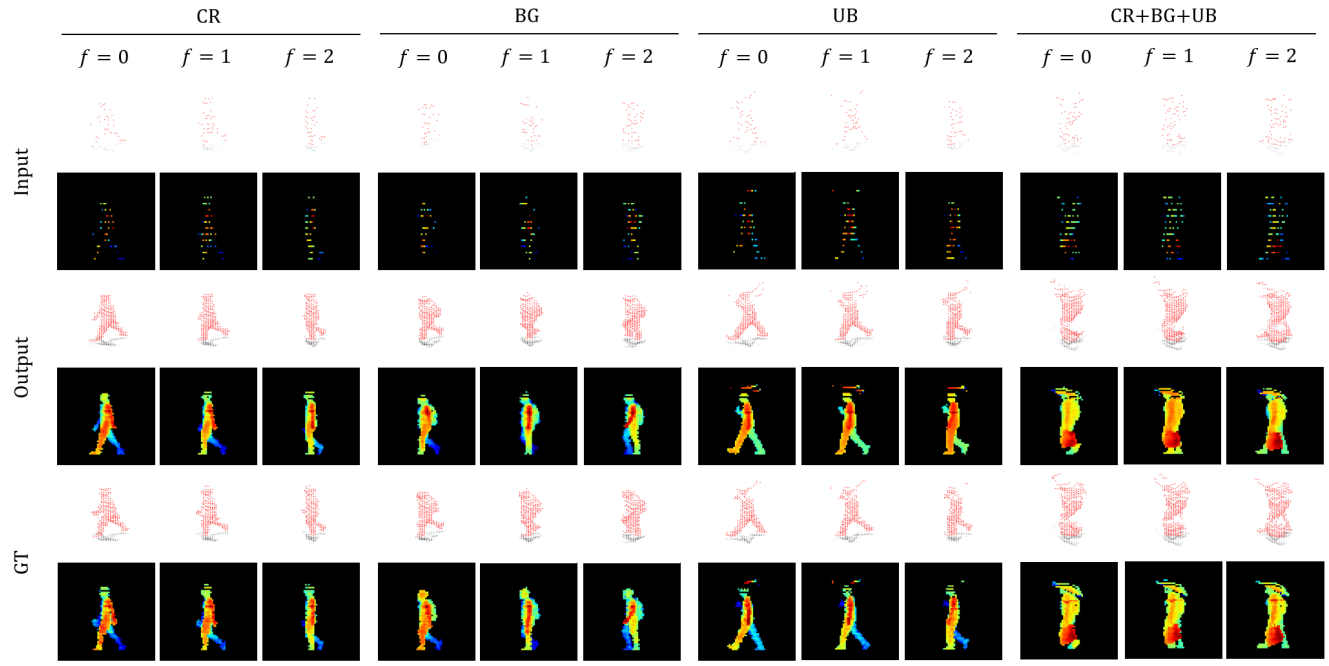

Fig. 5. Upsampled results using our model from noise masks with $\mathbf{V}\times 3/4$ and $\mathbf{P}\times 3/6$. We showcase the samples for three gait variances: Carrying (*CR*), Bag (*BG*), and Umbrella (*UB*).

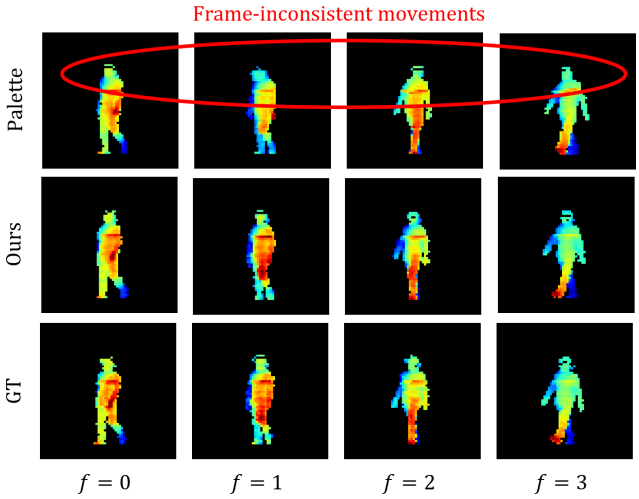

Fig. 6. Comparison between our model and Palette [16]. The results are sampled from noise masks with $\mathbf{V}\times 3/4$ and $\mathbf{P}\times 3/6$ (top two rows).

indicate how many times the neural networks are processed during sampling. For all noise mask combinations, it can be observed that overall performance generally improves as T increases, while remaining consistent even when T is reduced to 4.

E. Application

The identification results conducted on our dataset [2] using LidarGait [17] to evaluate the applicability of our model are shown in Table IV. In addition, the gait shapes according to the two different point cloud projections as illustrated in Fig. 8. In Table IV, we observed that our upsampling method

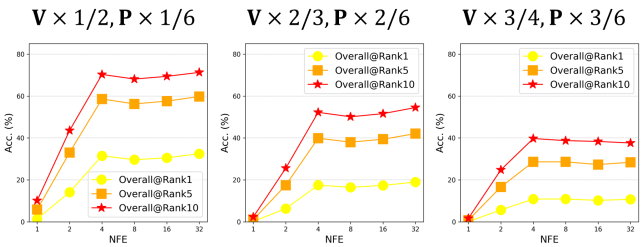

Fig. 7. Comparison of the number of function evaluations (NFE) for our model by sweeping T across $\{1, 2, 4, 8, 16, 32\}$.

and training strategy significantly contribute to performance improvement, even for real-world scenarios. Interestingly, the highest performance gain was achieved when both the probe set and the gallery set were fully restored.

TABLE IV
IDENTIFICATION RESULTS ON THE REAL-WORLD DATASET [2].

Upsampling				Overall	
Method	Gallery (10 m)	Probe (20 m)	Projection	Rank1 ↑	Rank5 ↑
			Spher.	5.51	25.98
			Ortho.	7.07	30.80
Palette [16]		✓	Ortho.	19.57	56.25
	✓	✓	Ortho.	25.45	63.54
Ours		✓	Ortho.	21.28	60.94
	✓	✓	Ortho.	**25.97**	**66.82**

V. CONCLUSIONS

In this study, we introduced a diffusion-based upsampling method for LiDAR-based gait sequence data, addressing a distance-independent inpainting problem. Our model demonstrated significant performance improvements compared with

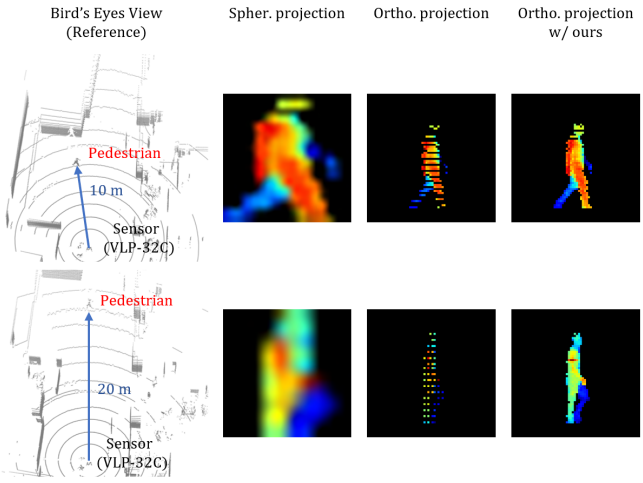

Fig. 8. Projection comparison on the real-world dataset [2] at two capture distances: spherical projection (Spher.) and our orthographic projection (Ortho.).

other methods in terms of both generation quality and gait recognition. Notably, our model is effective even for pedestrians with varying sensor resolutions or measurement distances in real-world scenarios. Future work will involve applying point-based identification models and investigating restoration for additional noise types, such as frame-drop noise and occlusions caused by obstacles.

References

[1] Jeongho Ahn, Kazuto Nakashima, Koki Yoshino, Yumi Iwashita, and Ryo Kurazume. 2v-gait: Gait recognition using 3d lidar robust to changes in walking direction and measurement distance. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pages 602–607, 2022.

[2] Jeongho Ahn, Kazuto Nakashima, Koki Yoshino, Yumi Iwashita, and Ryo Kurazume. Learning viewpoint-invariant features for lidar-based gait recognition. *IEEE Access*, 11:129749–129762, 2023.

[3] Csaba Benedek, Bence Gálai, Balá'zs Nagy, and Zsolt Jankó. Lidar-based gait analysis and activity recognition in a 4d surveillance system. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)*, 28(1):101–113, 2018.

[4] Hanqing Chao, Kun Wang, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Cross-view gait recognition through utilizing gait as a deep set. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44(7):3467–3478, 2022.

[5] Chao Fan, Junhao Liang, Chuanfu Shen, Saihui Hou, Yongzhen Huang, and Shiqi Yu. Opengait: Revisiting gait recognition towards better practicality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9707–9716, June 2023.

[6] 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 (NeurIPS)*, volume 27. Curran Associates, Inc., 2014.

[7] J. Han and Bir Bhanu. Individual recognition using gait energy image. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 28(2):316–322, 2006.

[8] Sander Elias Magnussen Helgesen, Kazuto Nakashima, Jim Tørresen, and Ryo Kurazume. Fast lidar upsampling using conditional diffusion models. In *IEEE International Conference on Robot & Human Interactive Communication (ROMAN)*, 2024.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 8633–8646, 2022.

[11] Diederik P. Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[12] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022.

[13] Kazuto Nakashima and Ryo Kurazume. Lidar data synthesis with denoising diffusion probabilistic models. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 14724–14731, 2024.

[14] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 July 2021.

[15] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi. Image super-resolution via iterative refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1867–1876, 2022.

[16] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10693–10703, 2022.

[17] Chuanfu Shen, Chao Fan, Wei Wu, Rui Wang, George Q. Huang, and Shiqi Yu. Lidargait: Benchmarking 3d gait recognition with point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1054–1063, June 2023.

[18] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *arXiv preprint arXiv:1503.03585*, 2015.

[19] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 12438–12448. Curran Associates, Inc., 2020.

[20] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[21] Torben Teepe, Ali Khan, Johannes Gilg, Fabian Herzog, Stefan Hörmann, and Gerhard Rigoll. Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In *IEEE International Conference on Image Processing (ICIP)*, pages 2314–2318, 2021.

[22] Hiroyuki Yamada, Jeongho Ahn, Oscar Martinez Mozos, Yumi Iwashita, and Ryo Kurazume. Gait-based person identification using 3d lidar and long short-term memory deep networks. *Advanced Robotics*, 34(18):1201–1211, 2020.

[23] Jinkai Zheng, Xinchen Liu, Wu Liu, Lingxiao He, Chenggang Yan, and Tao Mei. Gait recognition in the wild with dense 3d representations and a benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20196–20205, 2022.

[24] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point cloud. In *European Conference on Computer Vision (ECCV)*, 2022.