KYUSHU UNIVERSITY

Jet Propulsion Laboratory
California Institute of Technology

MUNICH 2025

Gait Sequence Upsampling using Diffusion Models for Single LiDAR sensors

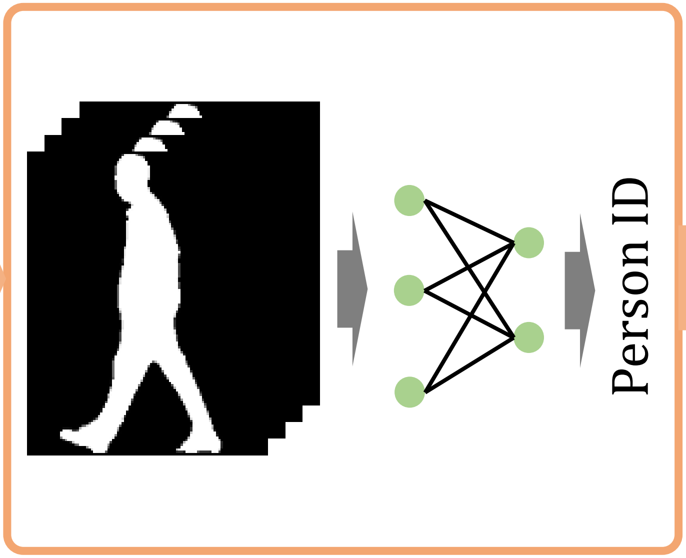
Jeongho Ahn[1], Kazuto Nakashima[1], Koki Yoshino[1], Yumi Iwashita[2] and Ryo Kurazume[1]
[1]Kyushu University
[2]JPL/NASA

Jan 22, 2025

Introduction / Gait Recognition

- **Biometric technology** that identifies people based on their **walking patterns**
- **Operate from a distance** without user's cooperation or physical contact

Measurement of pedestrian data using a visual device



ID matching with the database

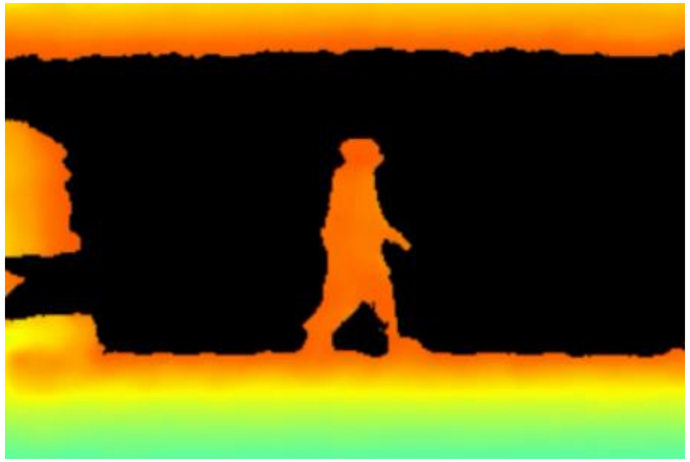
Person identification based on gait analysis [Fan+, CVPR'23]

Introduction / 3D LiDAR

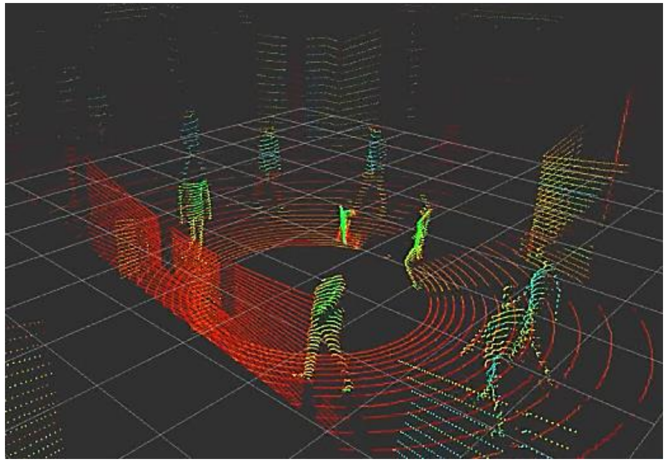
- Visualization comparison:

RGB camera

Depth camera



LiDAR sensor



Resolution — High

Field-of-View — Wide

Illumination — Robust

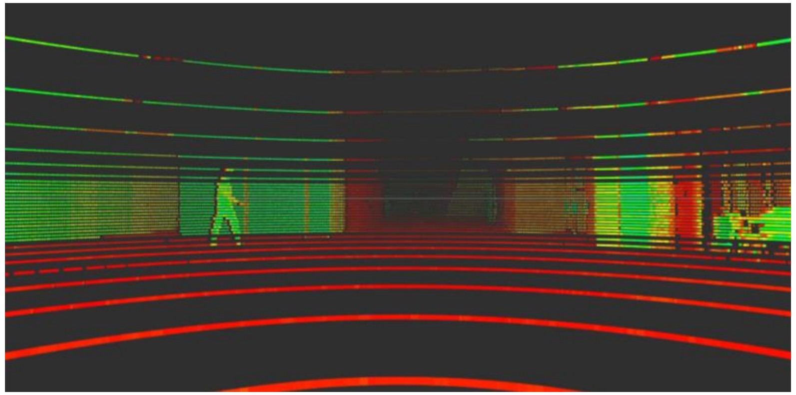
→ *Well-suited for **outdoor criminal investigations** or **security systems!***

Introduction / Motivation

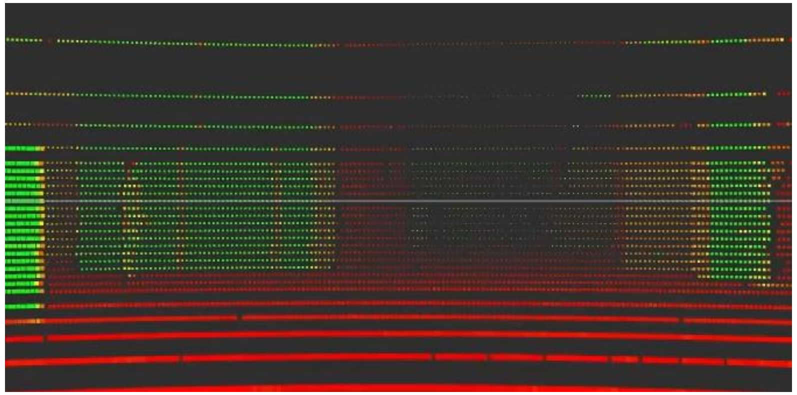
- Changes in **resolution/sparsity** based on **distances**

RGB camera
(reference)

LiDAR sensor (Velodyne VLP-32C)



10 m



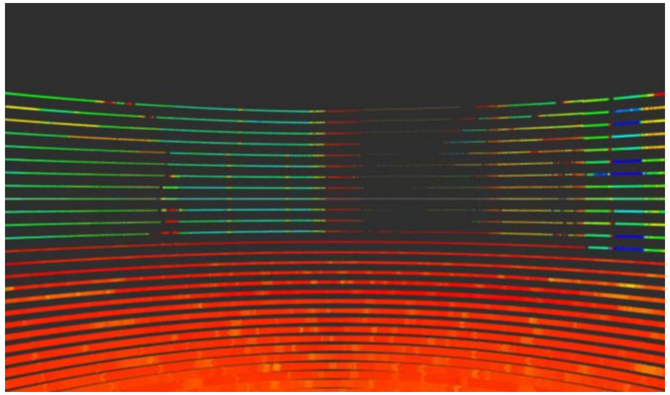
20 m

30 m

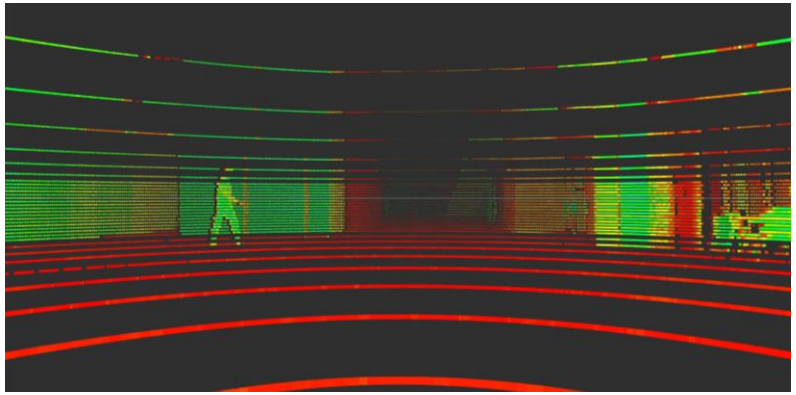
Introduction / Motivation

- Changes in **resolution/sparsity** based on **LiDAR sensor's emitting pattern (specification)**

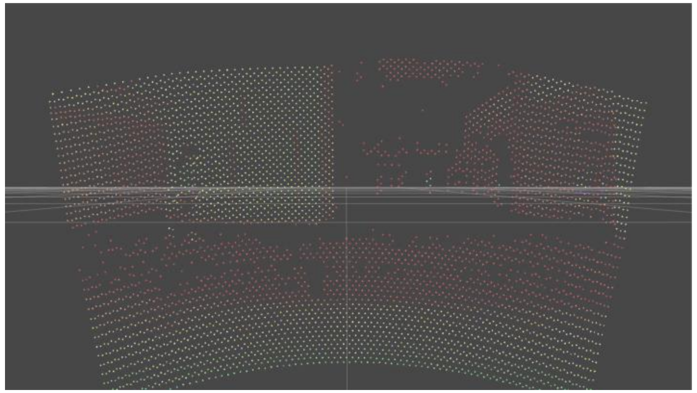
Velodyne HDL-32E



Velodyne VLP-32C



Pioneer SSL-S01



Dist: 10 m

Introduction / Motivation

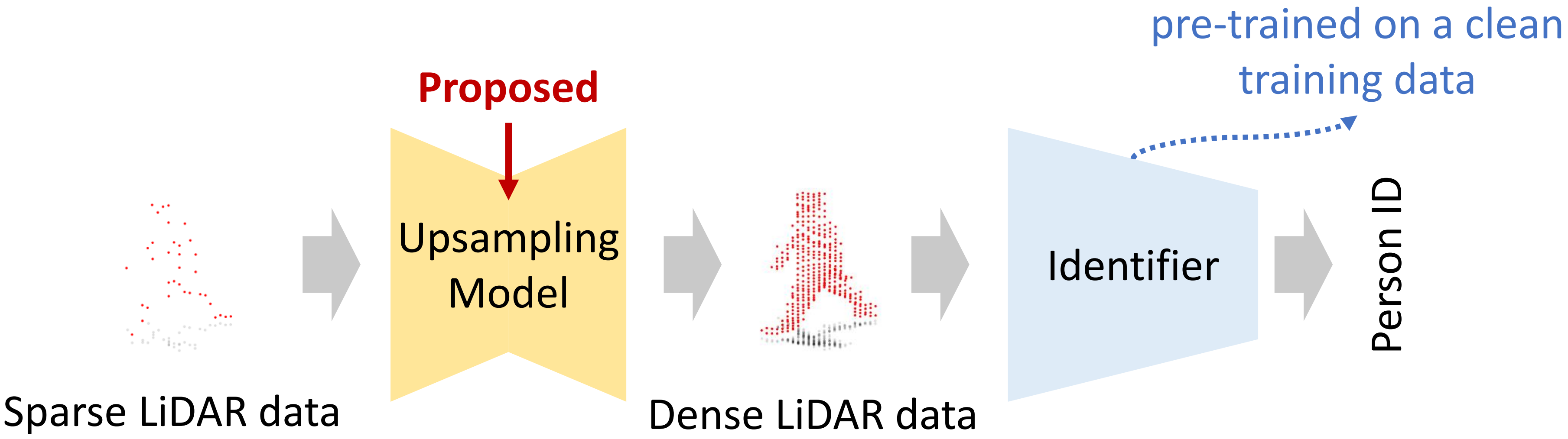
- Challenges:
  - Sparsity of LiDAR data is heavily influenced by **measurement distance** and **har dware specifications**
  - Collecting datasets for all **distances** and **sensor types** is practically impossible

→ *Necessary to reconstruct the* ***underlying/complete pedestrian shapes*** *from* ***sparse data****!*

Introduction / Goal

- Goals:
  - Develope **a gait sequence upsampling model** for sparse pedestrian data
  - Enhance **the generalization capability** of existing/future identification models

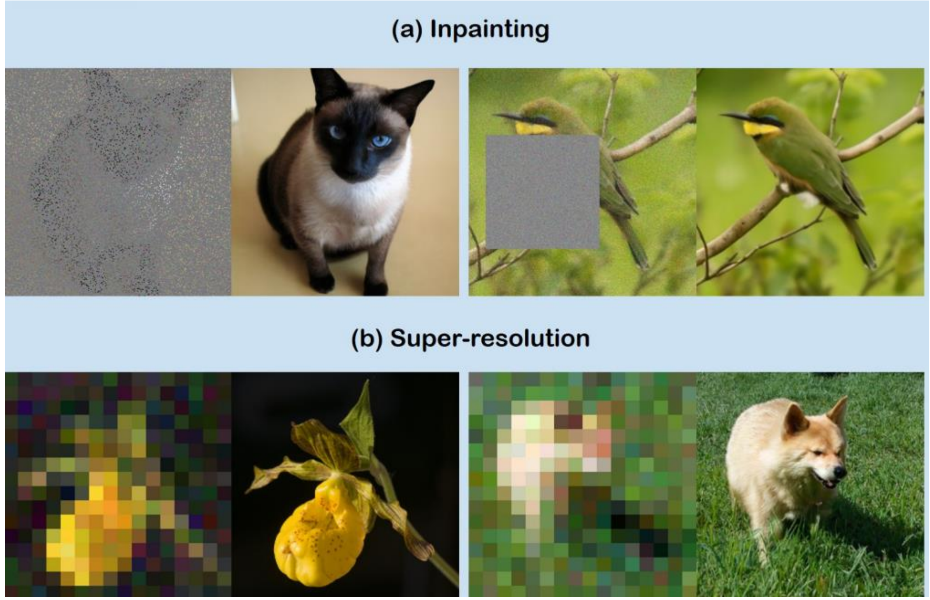
- Approches:
  - Employ **a video-based diffusion model**
  - Utilize **a distance-independent inpainting** strategy



Related Work

- Typical signal/image restoration:

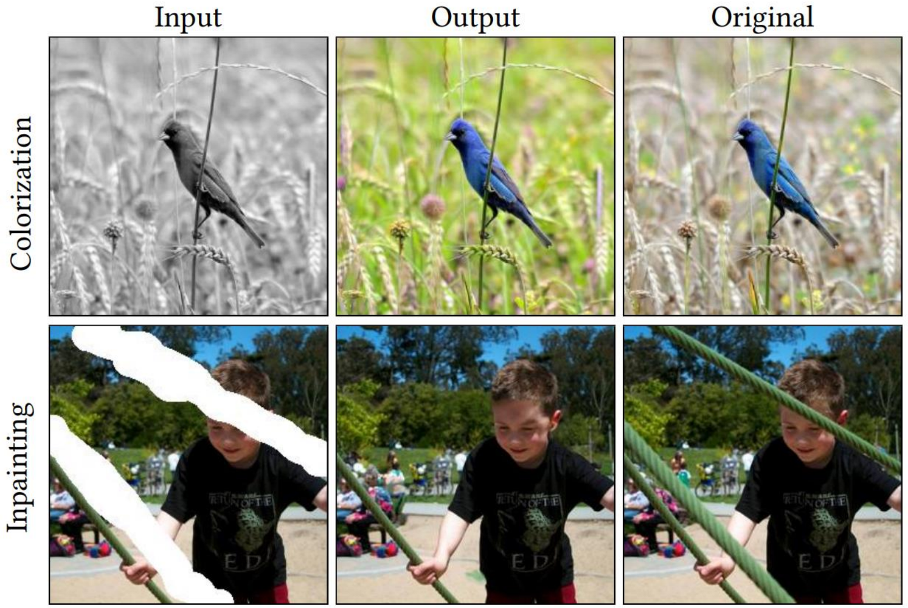
Task-agnostic approaches



Chung+, ICLR'23

- Learn the underlying data distribution using Bayes' rule
- Tend to worse than task-specific approaches

Task-specific approaches

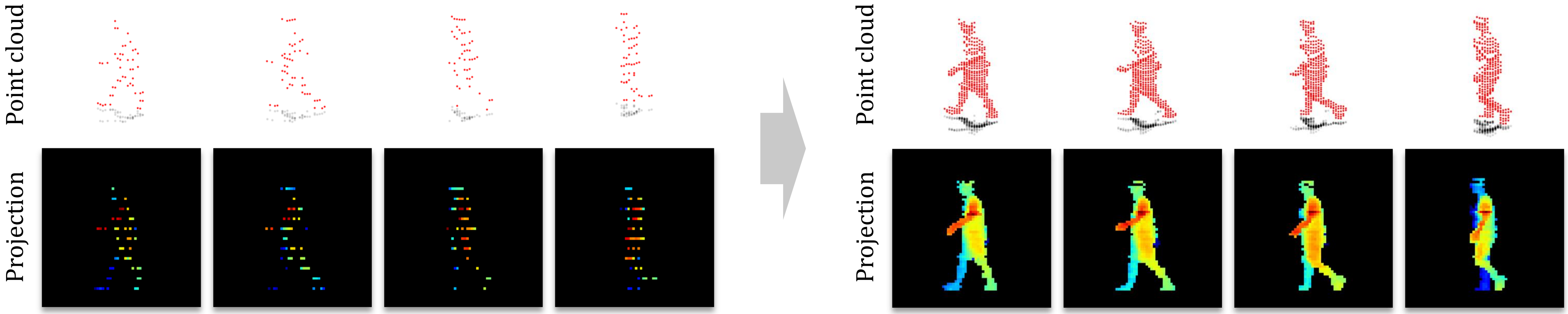


Saharia+, CVPR'22

- Conditional diffusion strategy
- Achieves superior performance across various multi-tasks

Method

- Overview



Input: sparse LiDAR gait sequences

Our results

Method / Problem Statement

- In **orthographic projection**, missing points in gait shapes can be addressed as **distance-independent inpainting problem**

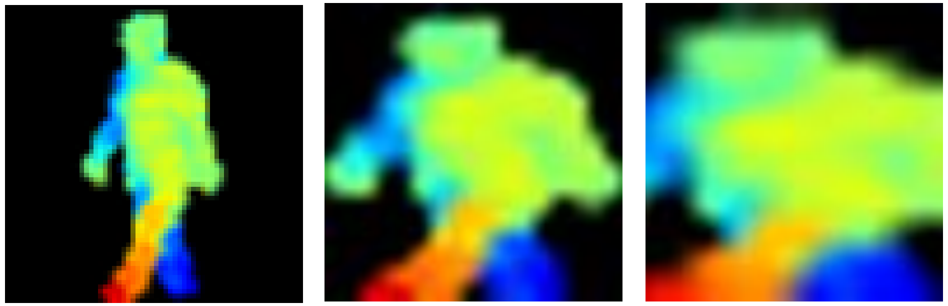
Degradation noise mask

Incomplete gait video $\mathbf{y} = H\mathbf{x}_0 + \mathbf{z}$ Gaussian noise

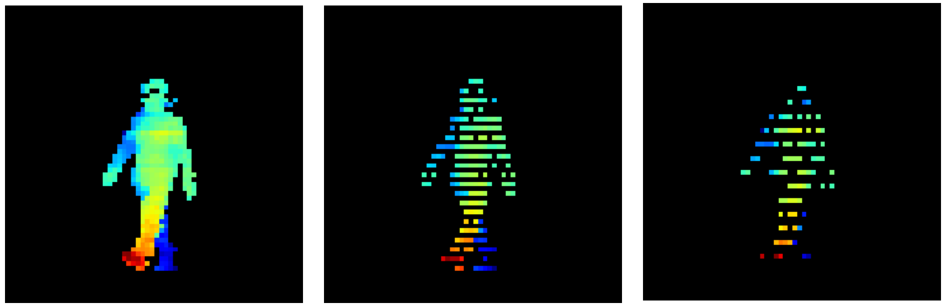
Complete gait video

$\mathbf{y}$ H $\mathbf{x}_0$

Spherical projection



Orthographic projection

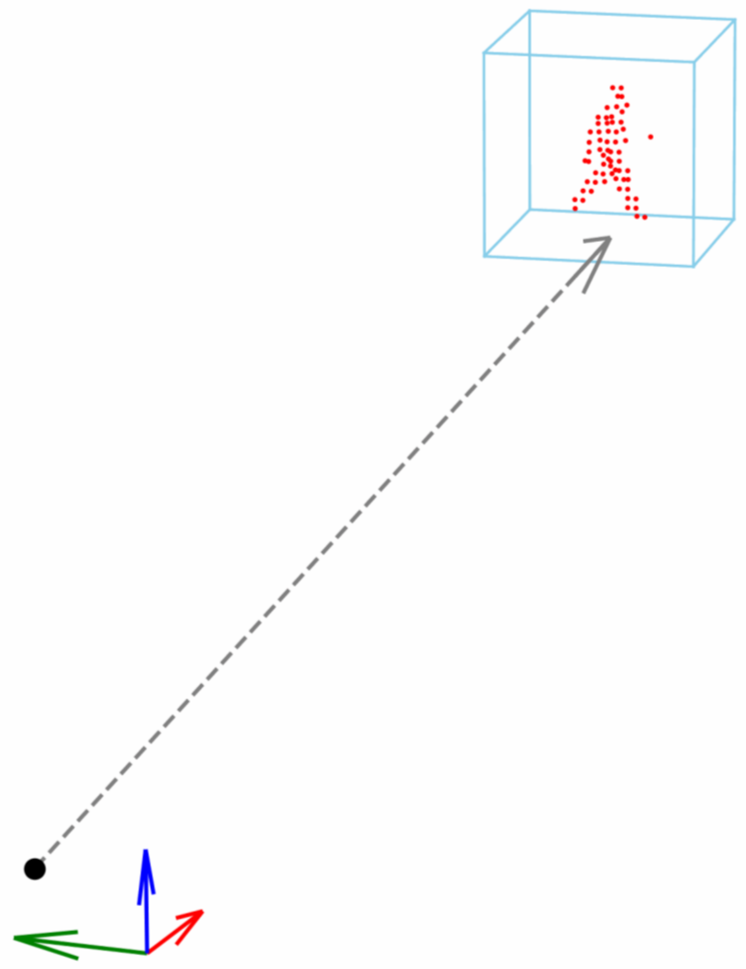


Method / Projection

- Transform a raw pedestrian point cloud sequenc e $\mathbf{P} \in \mathbb{R}^{F\times N\times C}$ into a depth video $\mathbf{y} \in \mathbb{R}^{F\times 1\times H\times W}$ from the **sensor-view**

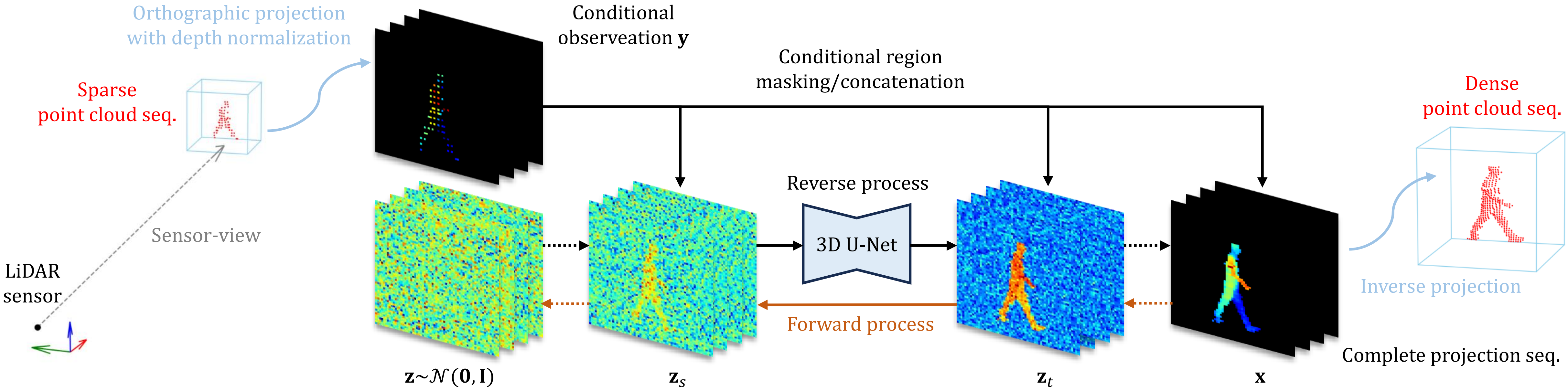
- Obtain the rotated point cloud sequence $\widehat{\mathbf{P}} \in \mathbb{R}^{F\times N\times C}$ with a directional angle $\theta_{\text{sensor},f}$:
  - $\theta_{\text{sensor},f} = \arctan(c_{f,y}, c_{f,x})$
  - $\hat{\mathbf{p}}_{f,n} = \left(\mathbf{P}_{f,n} - \mathbf{c}_f\right) \cdot \mathbf{R}_z(\theta_{\text{sensor},f} + \pi)$

- Project $\widehat{\mathbf{P}}$ onto the xz–plane



Method / Network

- Overall of the upsampling pipline

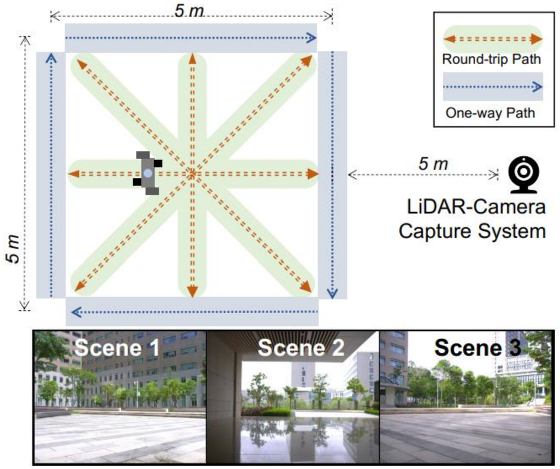


- Extended from Palette [Saharia+, CVPR'22]
- Initialization: $\mathbf{z}_t \leftarrow \mathbf{m} \odot \mathbf{y} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{z}_t$
- Loss function: $\mathcal{L}_{T\to\infty} = \mathbb{E}_{\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{1}), t \sim \mathcal{U}(0,1)}[\|\hat{\epsilon}(\mathrm{concat}(\mathbf{y}, \mathbf{z}_t); \lambda_t) - \epsilon\|_2^2]$

Experiments / Implementation Details

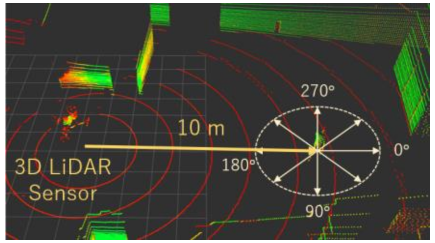
- Dataset comparison

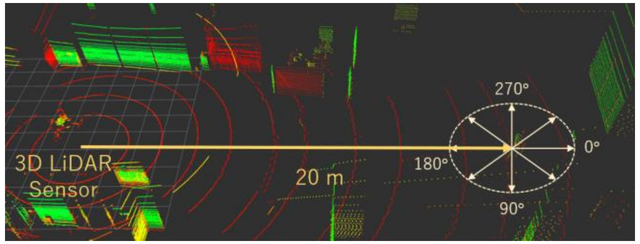
SUSTeck1K [Shen+, CVPR'23]



For generalization evaluation

Dataset used in Part I (2/2)





For practicality evaluation

Datasets	Sensors	Beams	V/H Resolutions	Subjects	Views	Distances
SUSTeck1K [56]	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

Experiments / Implementation Details

- **Noise masks** used for training and testing in the **generalization evaluation**

Pepper noise ($\mathbf{P}$)			Vertical lines ($\mathbf{V}$)		
$\times 1/6$	$\times 2/6$	$\times 3/6$	$\times 1/2$	$\times 2/3$	$\times 3/4$

- Simulate noise in the azimuth based on captured distances
- Represent the beam-level noise at the elevation of the LiDAR sensors

Experiments / Implementation Details

- **SUSTeck1K** dataset contains 1,050 subjects
  - Training set : **250 subjects**
  - Test set: **remaining 800 subjects**

- Learning settings:
  - Learning rate: 0.0003
  - input sequence length: 10 frames
  - Timesteps: 32

- Identifier (for the recognition task): **LidarGait** [Shen+, CVPR'23]
  - trained on the **clean training set** of SUSTeck1K

Experiments / Generative Evaluation

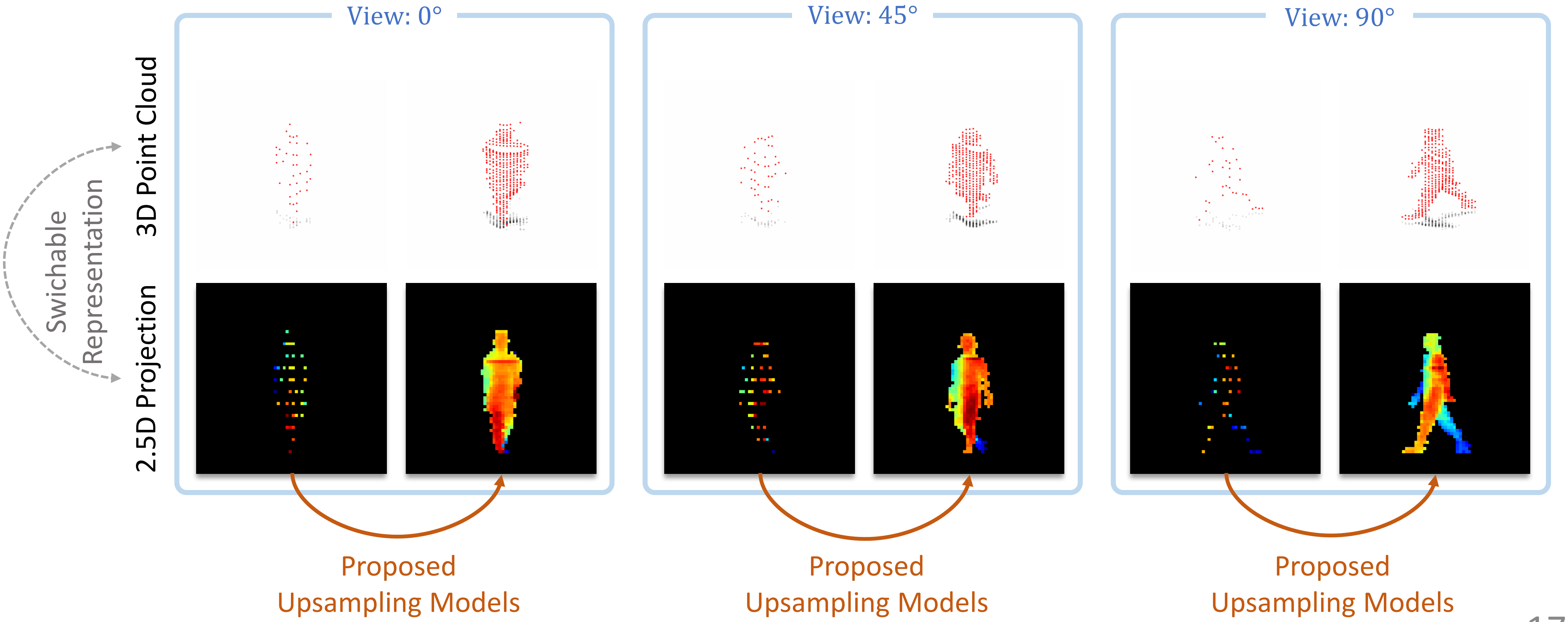
- Quantitative results:
  - Our model is Superior to all **linear interpolations** and **vanilla Palette** across three metrics

Table 4.2: Generative evaluation of the SUSTeck1K dataset with noise masks

Upsampling			Means (Test set)								
			V×1/2, **P**×1/6			**V**×2/3, **P**×2/6			**V**×3/4, **P**×3/6		
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 [52]	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**

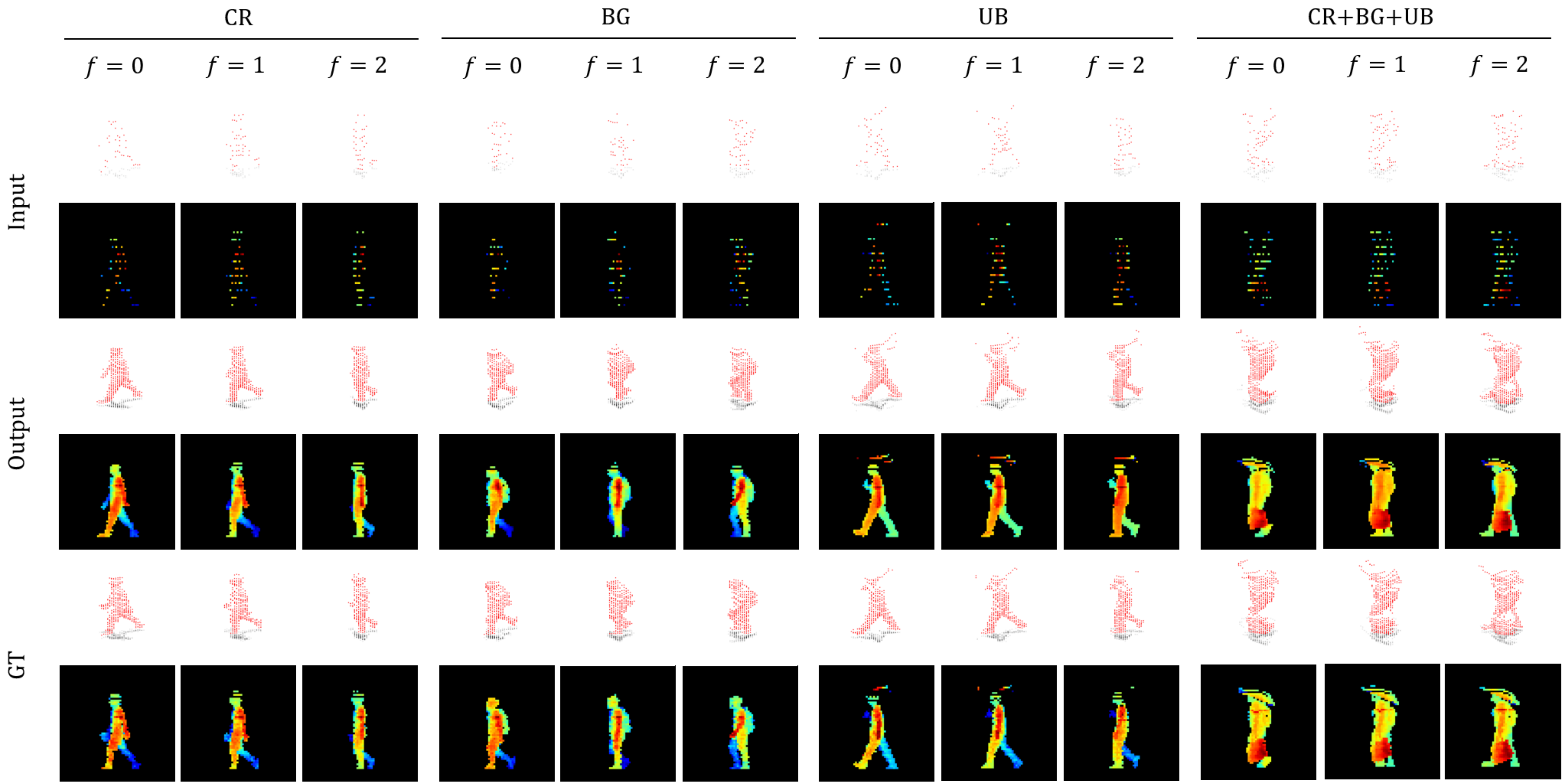
Experiments / Generative Evaluation

- Upsampled results using the proposed model **across three angles** on SUSTeck1K



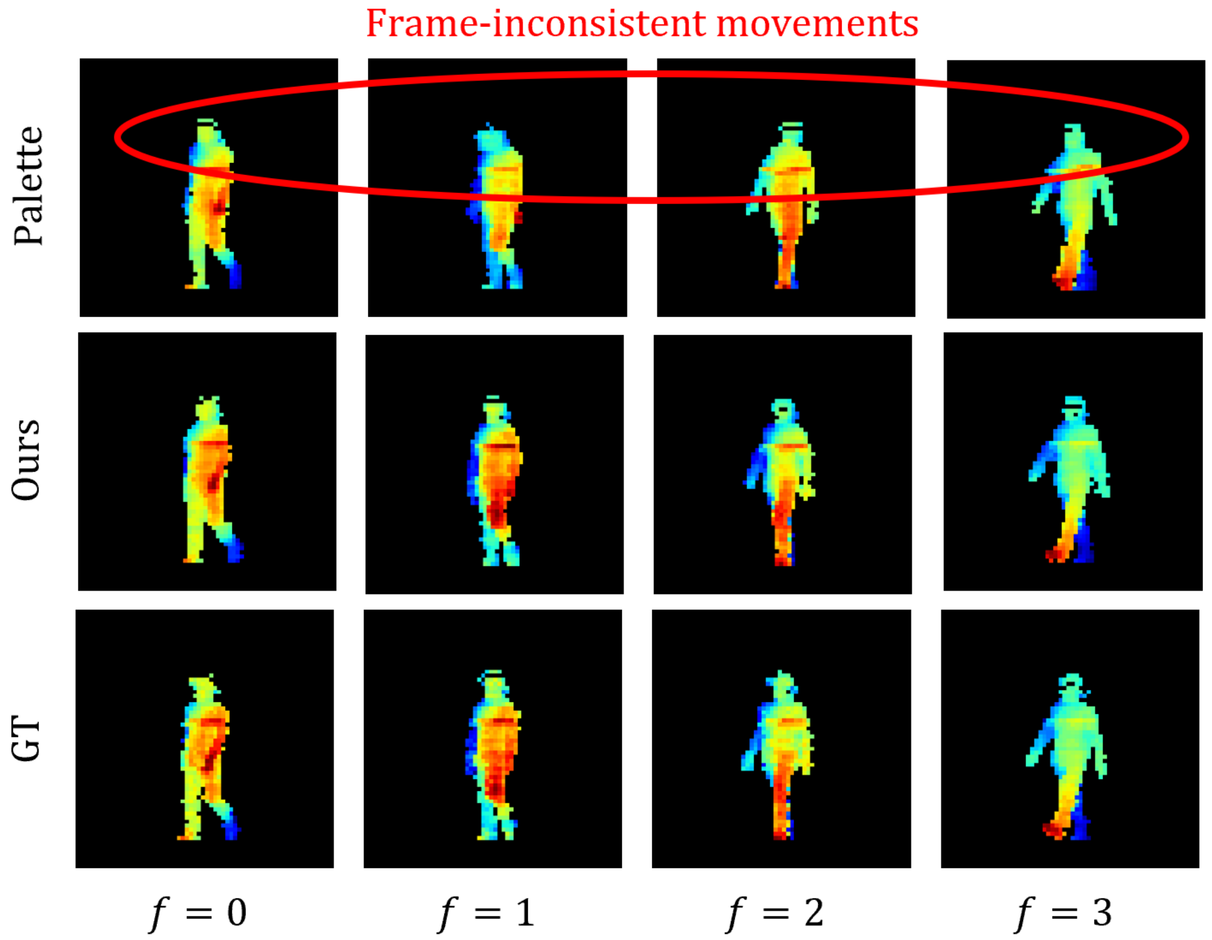
Experiments / Generative Evaluation

- Upsampled results **with various attributes** using the proposed model on SUSTeck1K



Experiments / Generative Evaluation

- Comparison between the proposed model and **vanilla Palette** [Saharia+, CVPR'22]:
  - The proposed model preserves **frame-consistency** more effectively



Experiments / Gait Recognition Task

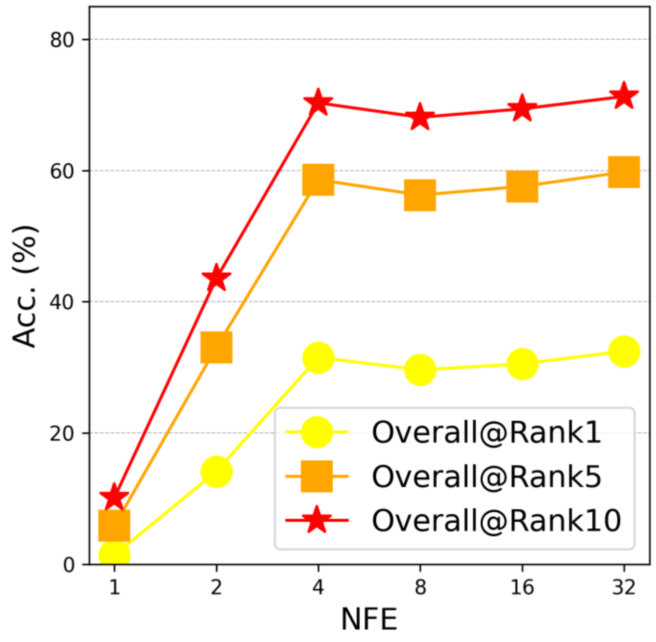
- Quantitative results:
  - As the noise masks become more severe, **the performance gap** between the proposed model and the original Palette **increases**

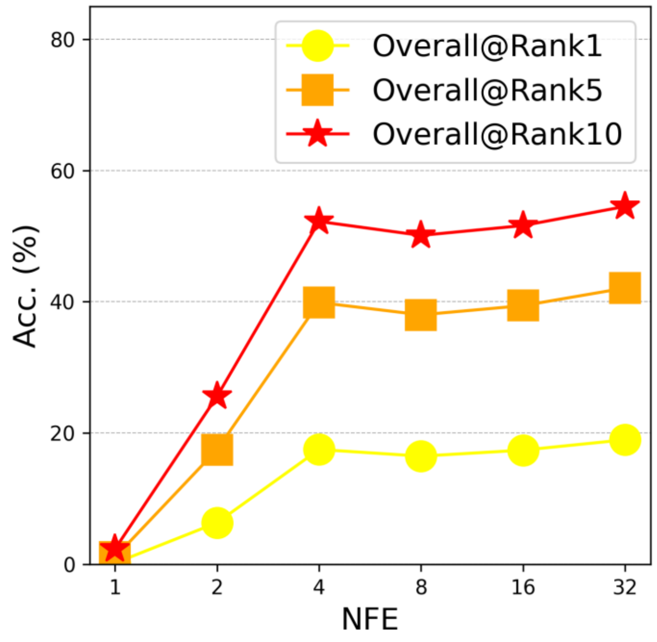
Table 4.3: Identification Evaluation using a LidarGait on SUSTeck1K dataset with noise masks

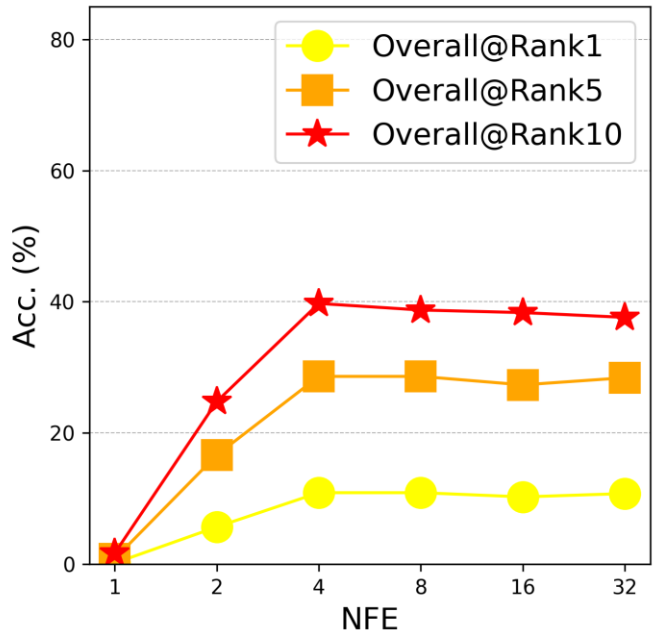
Upsampling			Means (Probe set)								
			V×1/2, **P**×1/6			**V**×2/3, **P**×2/6			**V**×3/4, **P**×3/6		
Approach	Method	Input Modality	Rank1 ↑	Rank5 ↑	Rank10 ↑	Rank1 ↑	Rank5 ↑	Rank10 ↑	Rank1 ↑	Rank5 ↑	Rank10 ↑
			1.40	5.85	10.13	0.18	1.08	2.34	0.15	0.82	1.68
Interpolation	Nearest-neighbor	Depth Image	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 [52]	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**

Experiments / Gait Recognition Task

- Comparison of the number of function evaluations (NFEs) for the proposed model







Experiments / Practicality

- Quantitative results:
  - Training set: **SUSTeck1K** with noise masks (with **128**-beam LiDAR sensosr)
  - Testing set: our collected dataset (with **32**-beam LiDAR sensosr)

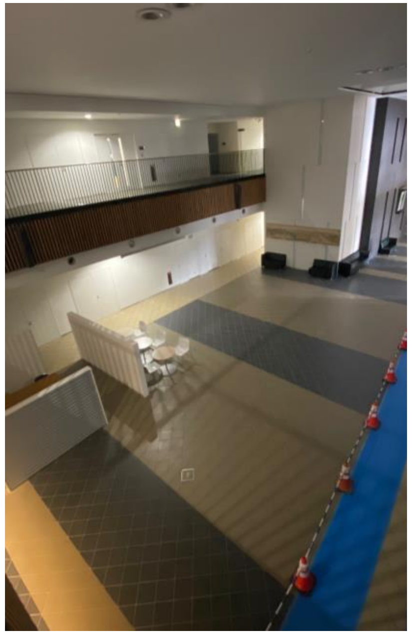
Table 4.4: Identification results on the real-world dataset [2].

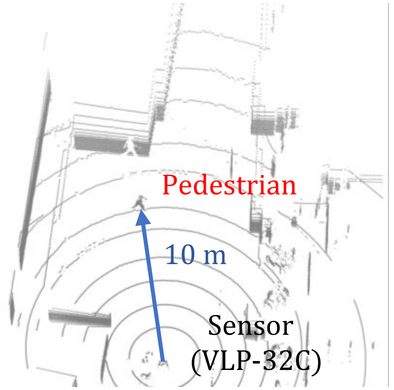
Upsampling				Overall	
Method	Gallery (10 m)	Probe (20 m)	Projection	Rank1 ↑	Rank5 ↑
			Spher.	5.51	25.98
			Ortho.	7.07	30.80
Palette [52]		✓	Ortho.	19.57	56.25
	✓	✓	Ortho.	25.45	63.54
Ours		✓	Ortho.	21.28	60.94
	✓	✓	Ortho.	**25.97**	**66.82**

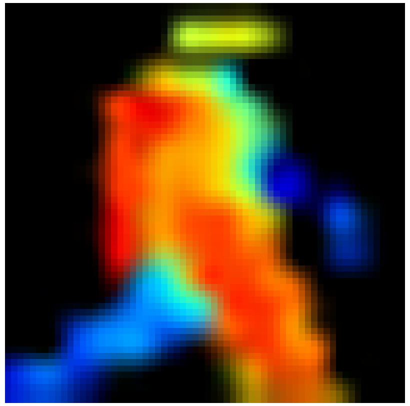
Experiments / Practicality

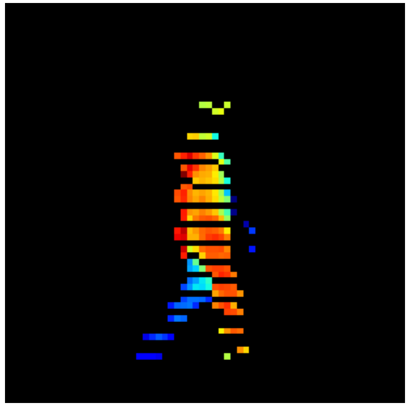
- Qualitative results

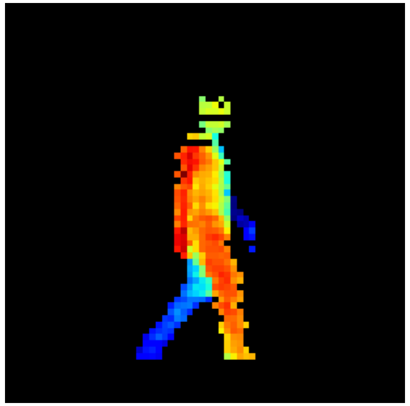
Environment (Reference) | Bird's Eyes View (Reference) | Spher. projection | Ortho. projection | **Ortho. projection w/ ours**

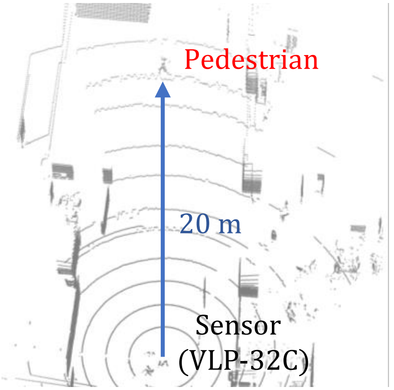


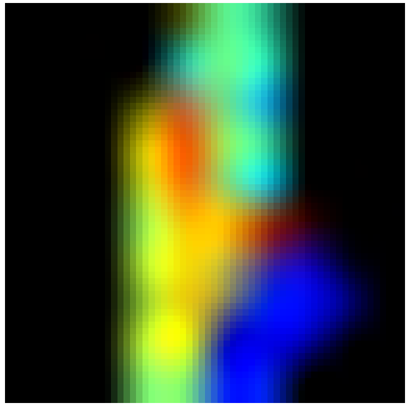


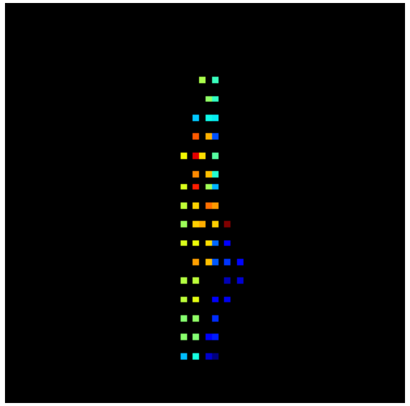


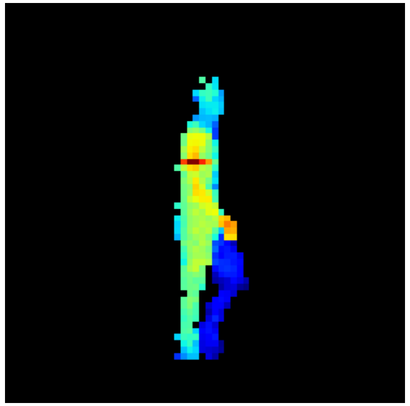






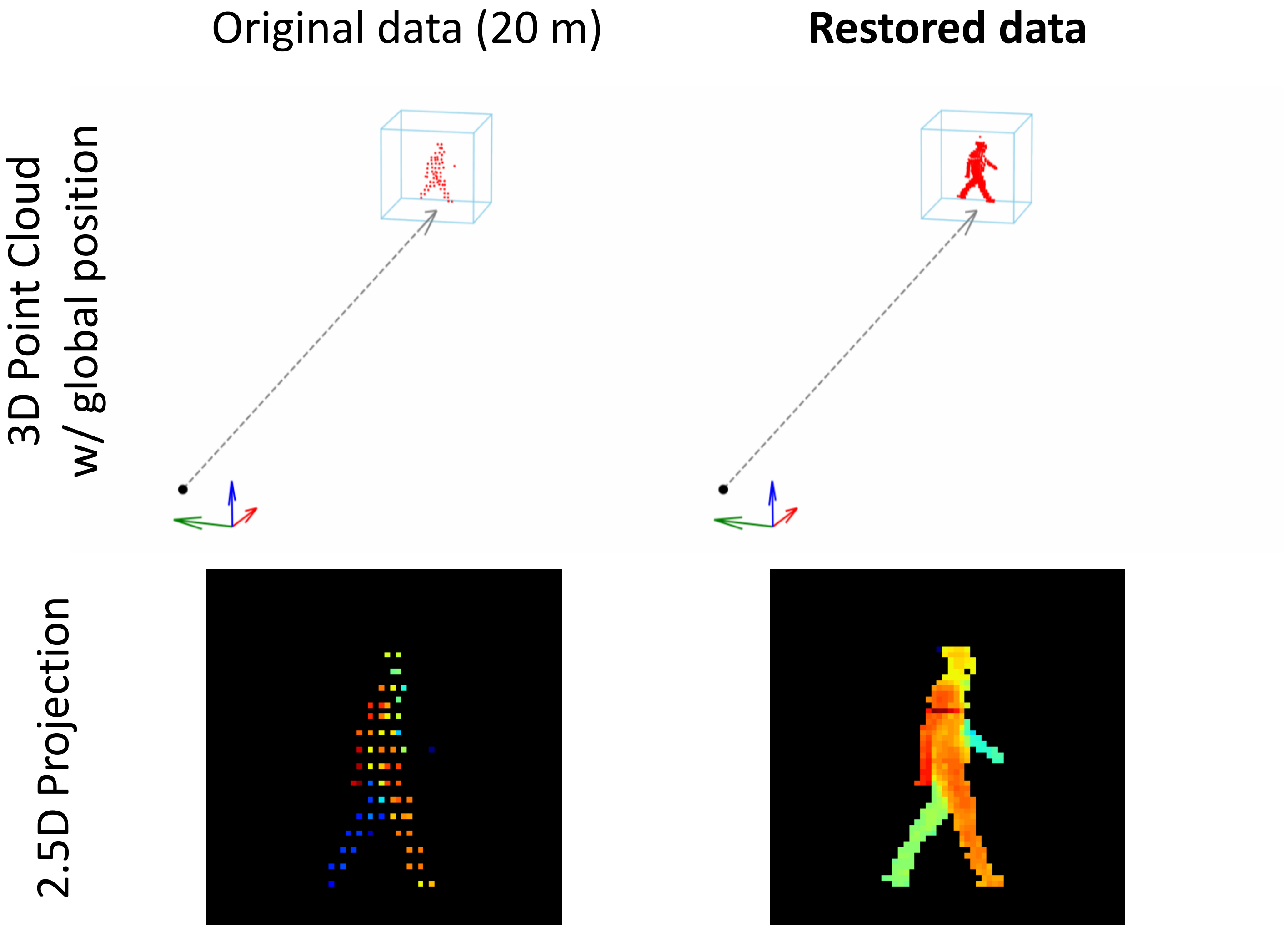






Experiments / Practicality

- Qualitative results



Summary

- Introduced an upsampling model for LiDAR-based gait sequence data to address a distance-independent inpainting problem
- Demonstrated significant improvements in terms of both generation quality and id entification performance
- Proved effectiveness even for varying sensor resolution or measurement distance i n real-world senarios

Thank you for your attention!