# 3D LiDAR-based Gait Analysis for Person Identification in Long-range Measurement Environments


**Jeongho Ahn**

Kurazume and Kawamura Lab., Kyushu University


PhD Dissertation Defense

Feb 18, 2025

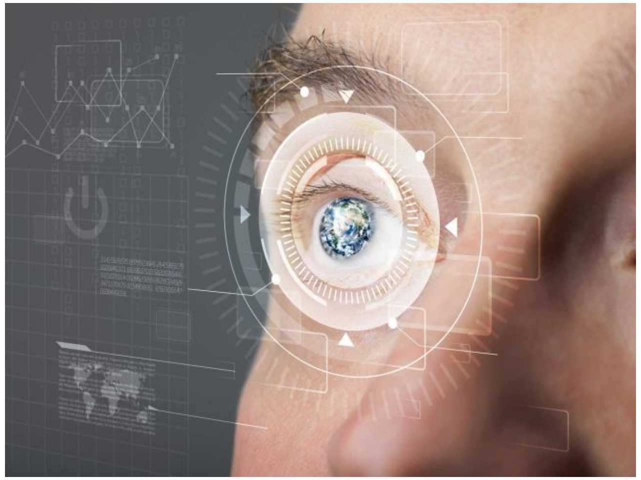
# Outline

- **Introduction**
- **Part 1: Development of gait recognition models using 3D LiDAR**
  - Identification modeling for range variations
  - Identification modeling through adaptive learning
- **Part 2: Development of gait upsampling models for 3D LiDAR**
  - Restoration modeling for gait sequence data
- **Conclusion**

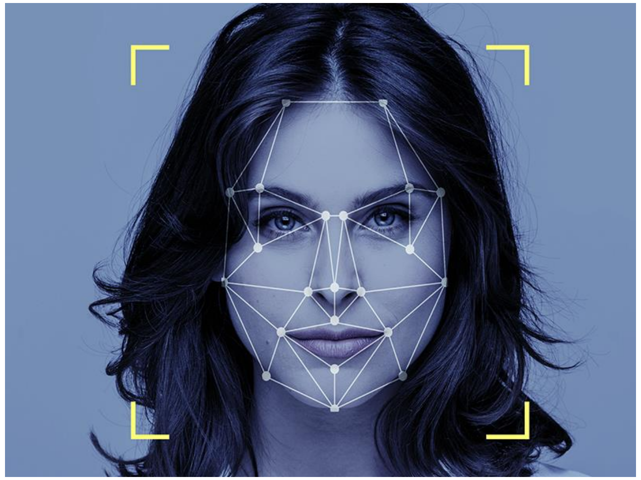
# Introduction / Person Identification

- Biometrics
  - Technologies using **physical characteristics** to identify individuals
  - Achieved **substandtial advancements** thanks to progress in **AI**

- Typical modalities



Iris

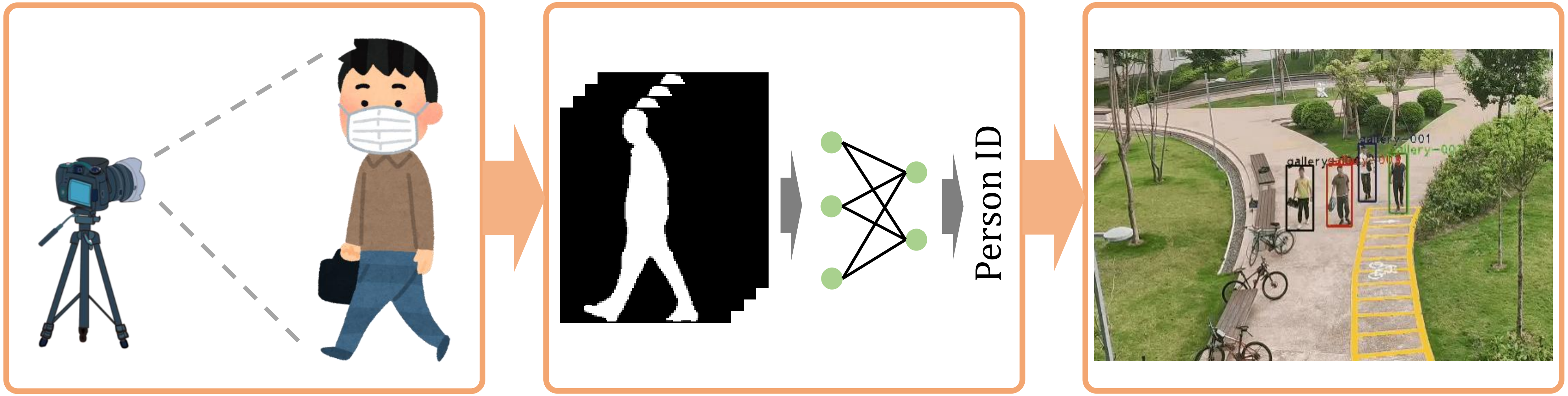


Face


Fingerprint

# Introduction / Gait Recognition

- Biometric technology that identifies people based on their **walking patterns**
- Operates 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 / Camera-based Identification

- Main device for gait recognition system so far: **RGB cameras**

**Pros**

- Ease of use (low cost)
- High spatial resolution

**Cons**

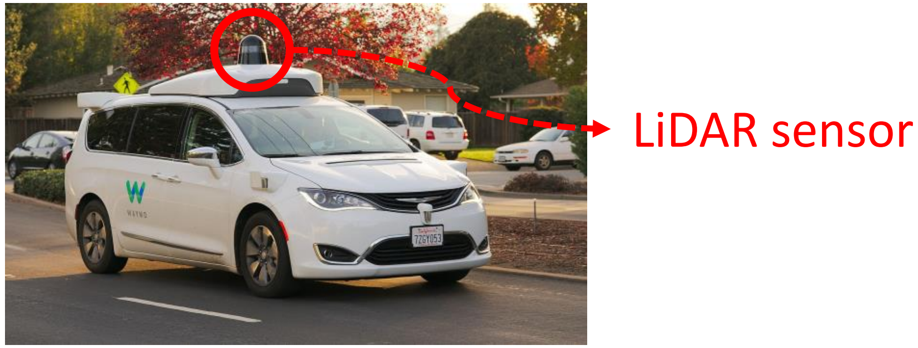
- Leak 3D geometry information
- Sensitive to lighting conditions
- Senstive to varying camera's height/angle

Night attribute
[Shen+, CVPR'23]

High-angle condition
[Zheng+, CVPR'22]

# Introduction / 3D LiDAR

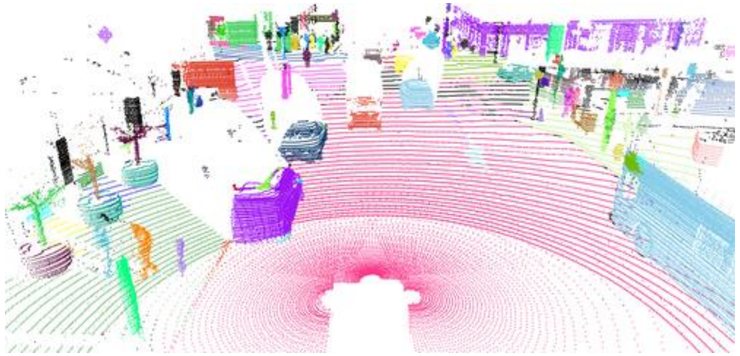
- Lighting Detection and Range (LiDAR)
  - 3D sensors scanning of surrounding environments



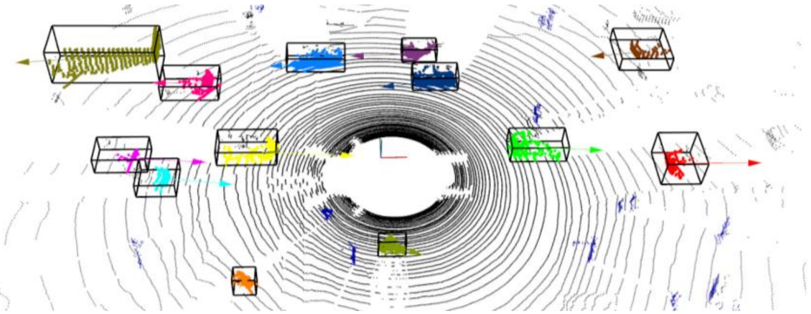

Self-driving texi (Waymo)

- Well-suited for outdoor applications

Semantic segmentation



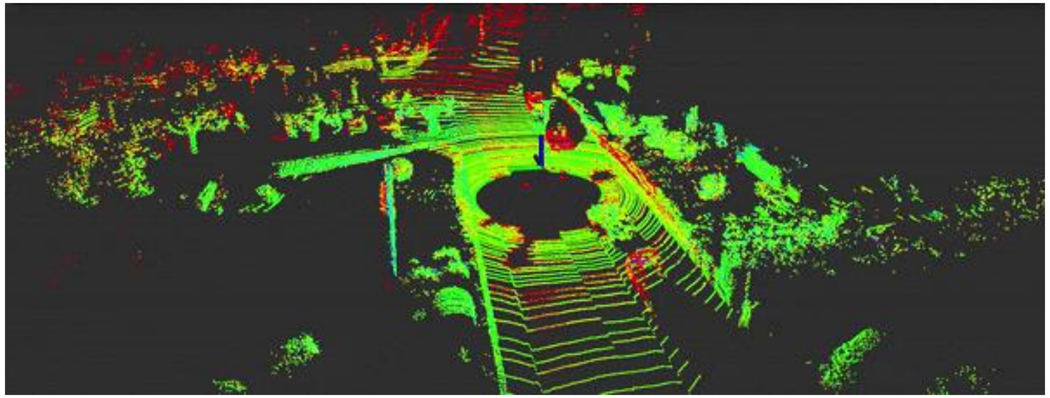
Object dection/tracking



# Introduction / 3D LiDAR

- LiDAR representation comparison

3D point clouds



- Three or more coordinates
- Raw geometric data
- Unordered nature
- Time-consuming computations

2D range images
(Spherical projection)

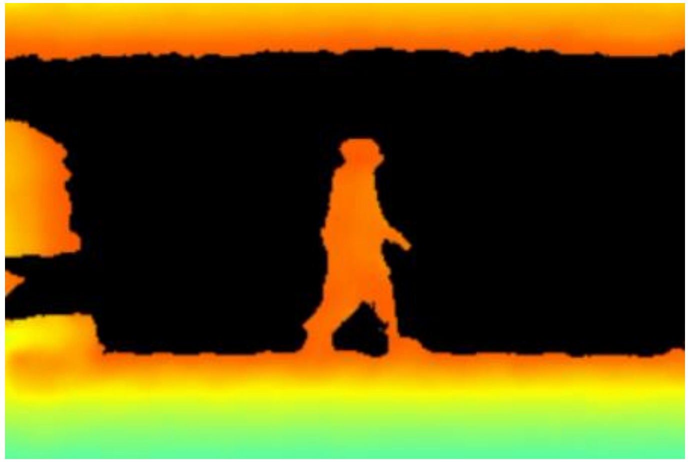
- 2D ordered formats
- Ease of use (more practical)
- Quantization artifacts

# Introduction / 3D LiDAR

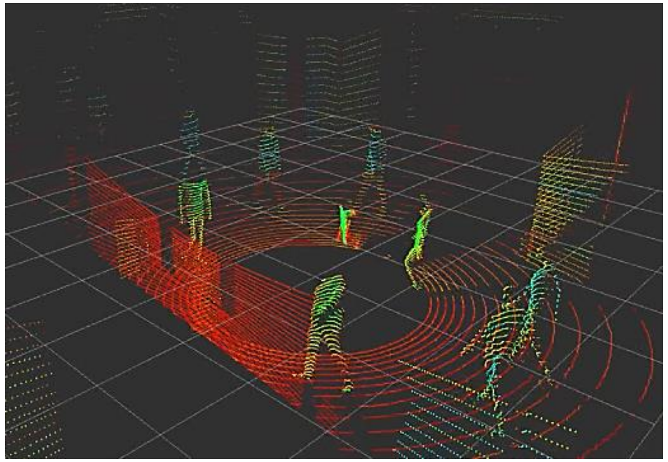
- Visualization comparison

RGB camera

Depth camera



LiDAR sensor



Resolution — High

Field-of-View — Wide

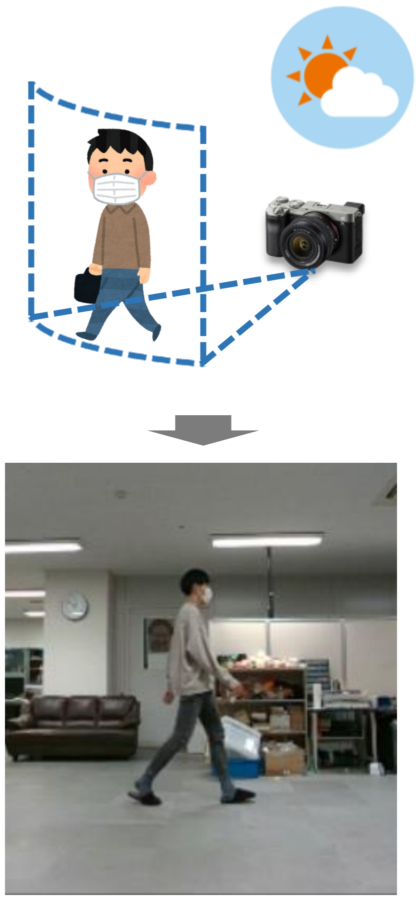
Illumination — Robust

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

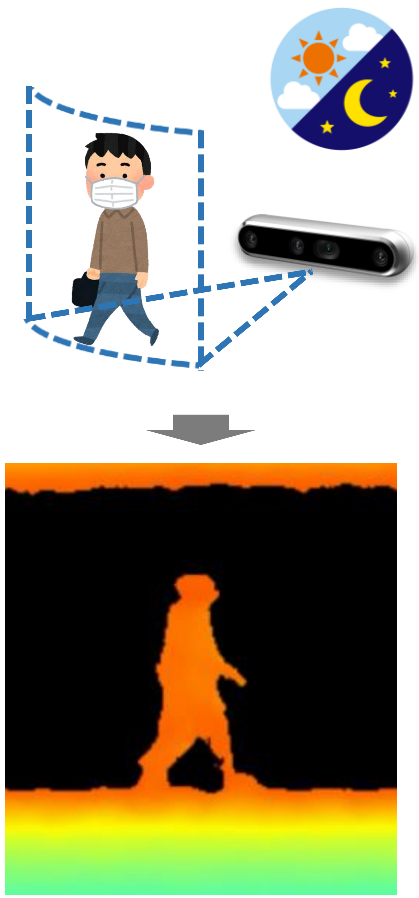
# Introduction / 3D LiDAR

- Visualization comparison

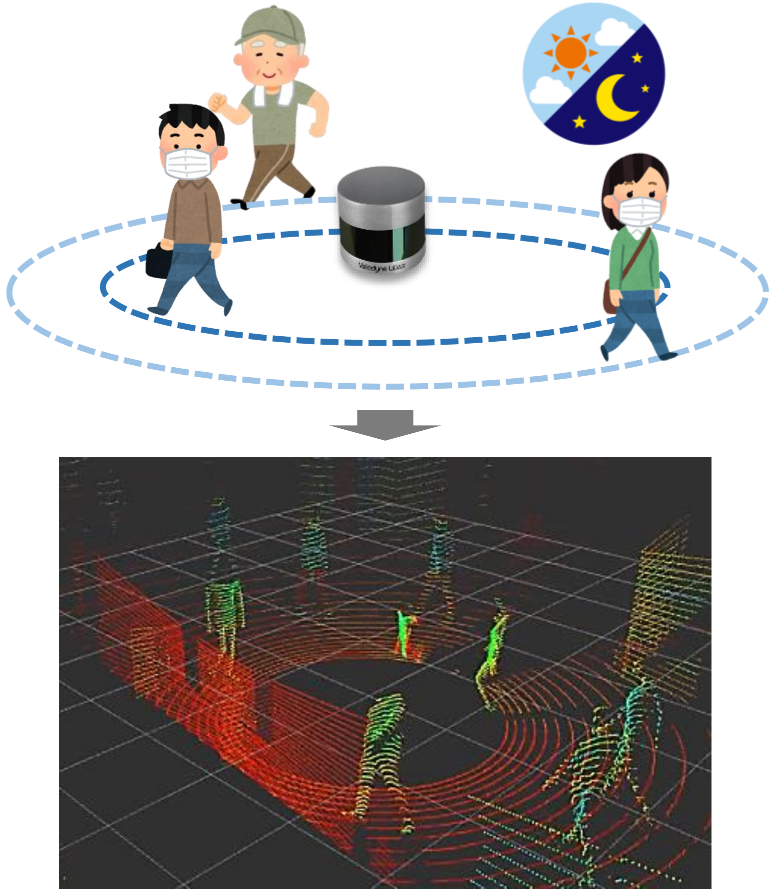
RGB camera



Depth camera



LiDAR sensor



# Introduction / LiDAR-based Gait Recognition

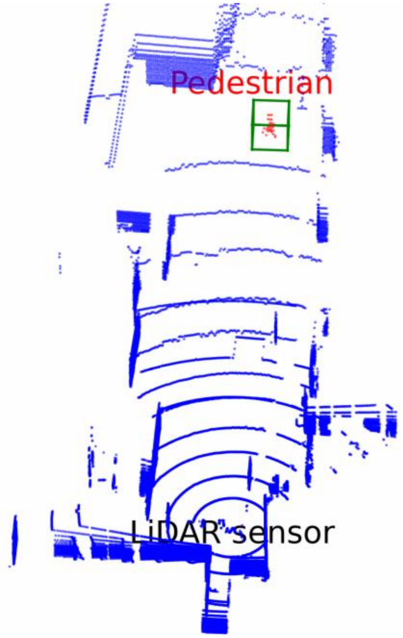
## Pros

- Wide range of directions/distances
- Robust to adverse weather
- Precise 3D geometry information

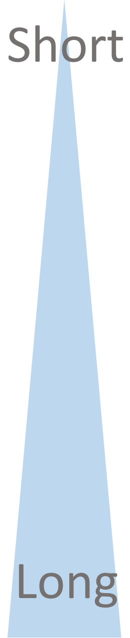
## Cons

- Sensitive to distances
- Poor spatial resolution (sparse data)

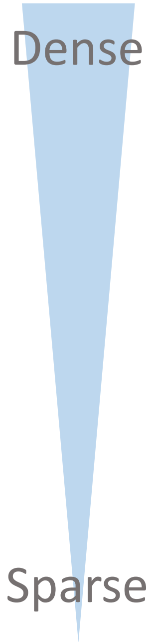
LiDAR data visualization




Distance




Sparsity




Gait data

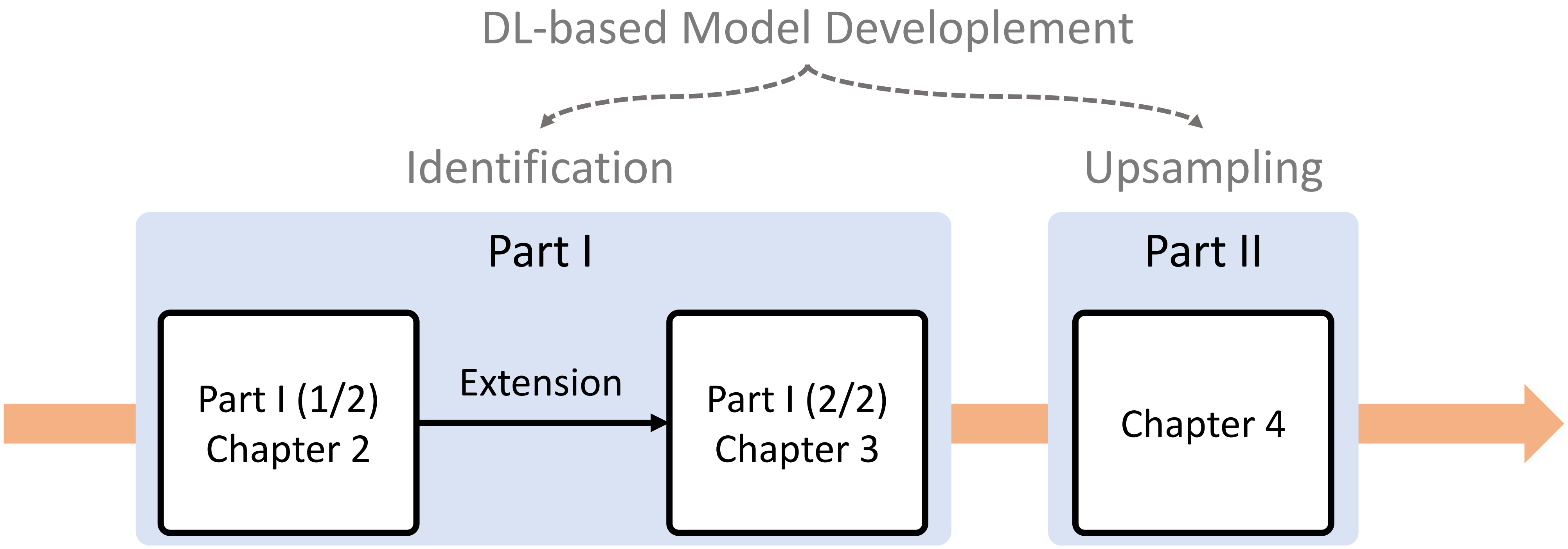
# Introduction / Goals

- Primary challenge:
  - Sparse pedestrian data caused by **long distances**

- Goal:
  - **Improve person identification performance** by using **deep learning techniques**

# Introduction / Goals

- Explored **in two aspects**
  - Part I: Development of **gait recognition models** using 3D LiDAR
  - Part II: Development of **gait upsampling models** for 3D LiDAR




Dessertation/presentation flow

# Part I (1/2): Development of Gait Recognition Models using 3D LiDAR

# Part I (1/2) / Motivation

- Applications using LiDAR-based person identification:

Security robots


- Operated 24h a day
- Nighttime surveillance system
- Less conspicuous than humans

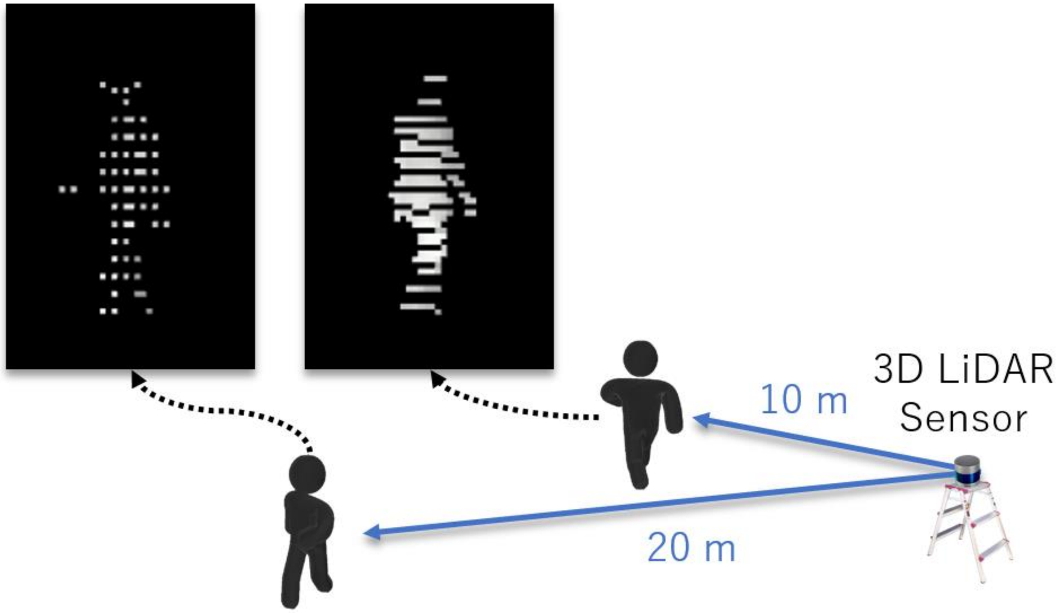
Autonomous vehicles

- Identify specific users
- Detect elderly people

# Part I (1/2) / Motivation

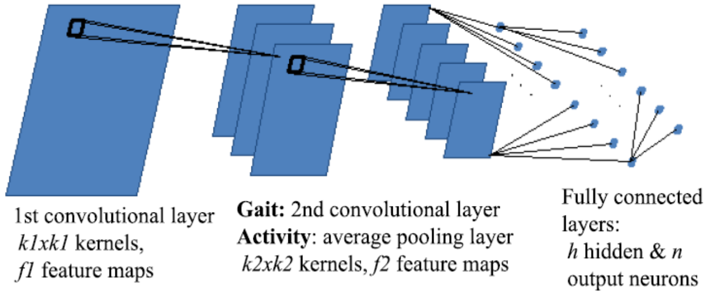
- Necessary to design **a robust identification model** for intra-subject changes:
  - Viewing angles
  - Measurement distances

- Invariant gait features under **these complex conditions**:
  - Two fixed viewpoints
  - Walking pace



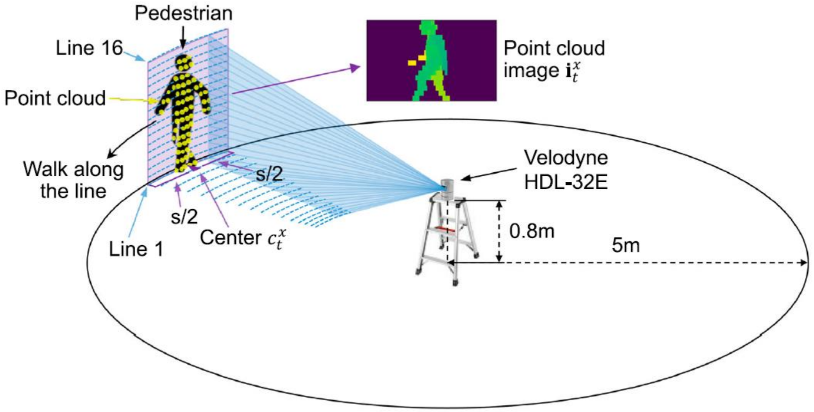
# Part I (1/2) / Related Work

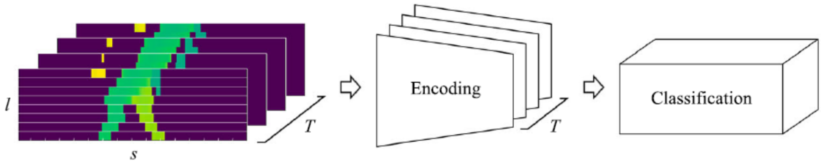
GEI-based identifier
[Benedek+, IEEE T-CSVT’18]




Difficult to extract the **dynamic feature** under **temporal changes**

Depth-based identifier
[Yamada+, Advance Robotics’20]

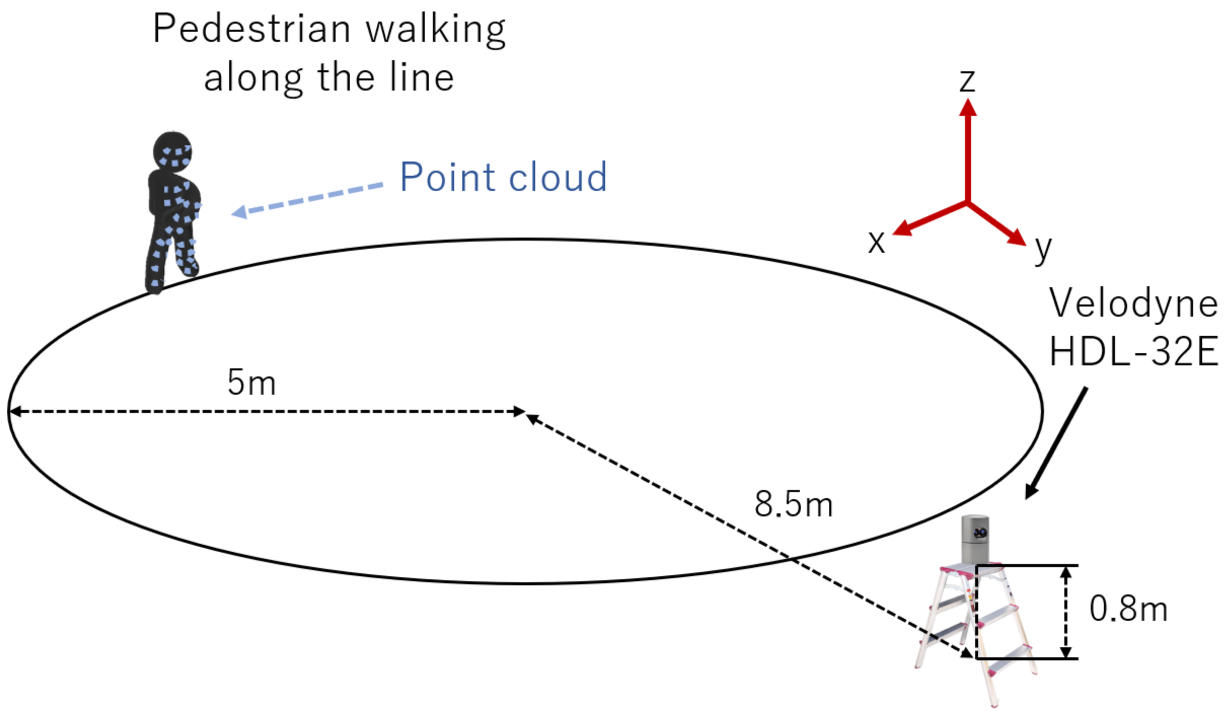


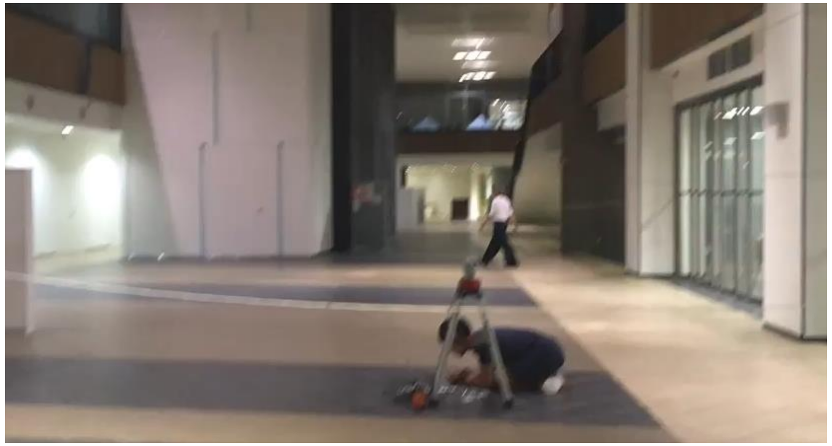




Performance degradation when the **distance/direction** is **not constant**

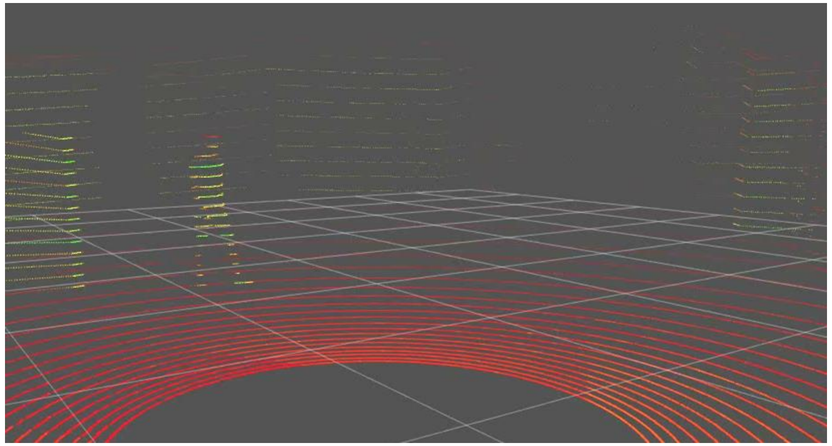
# Part I (1/2) / Dataset

- Captured using a **Velodyne HDL-32E**
- Collected gait sequence data from **31 subjects**



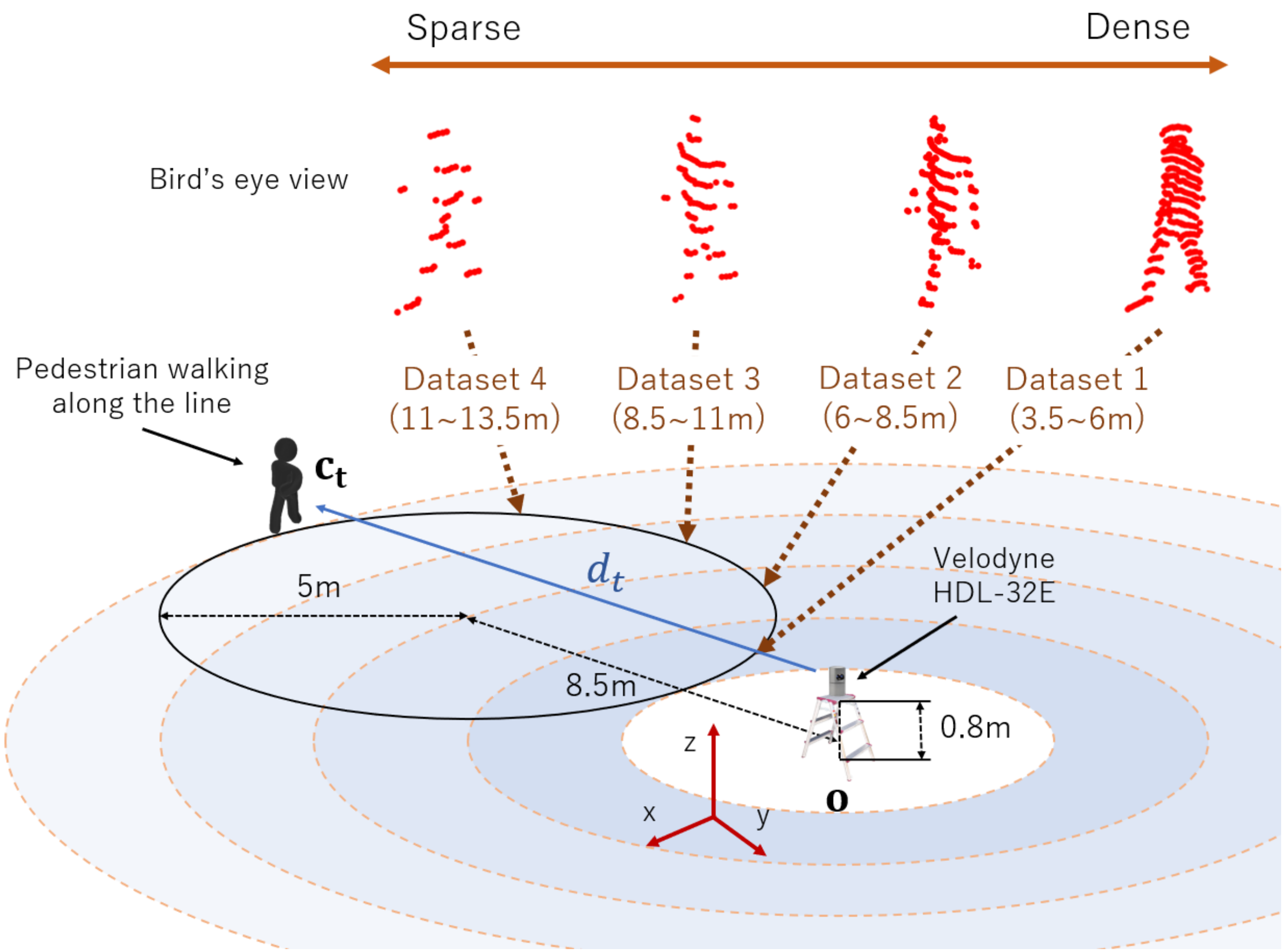
Data acquisition environment



LiDAR data visualization

# Part I (1/2) / Dataset

- Divided into **4 subsets** according to the distance $d_t$
  - Subset 1: 3.5-6 m
  - Subset 2: 6-8.5 m
  - Subset 3: 8.5-11 m
  - Subset 4: 11-13.5 m

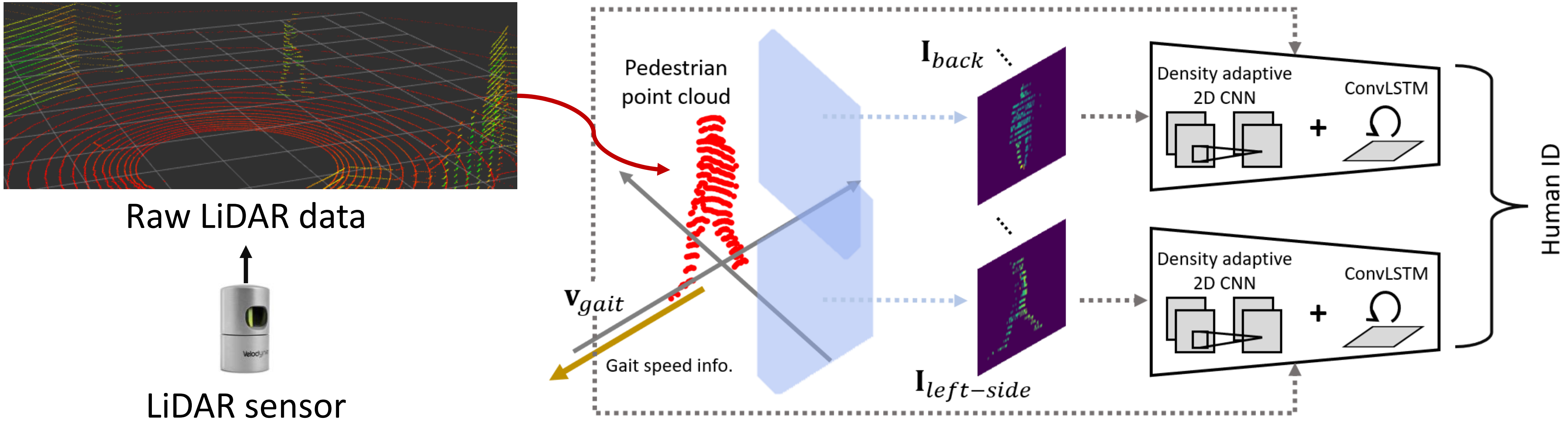



Contain the changes in the 360 walking direction and the distance **from 3.5 to 13.5 m**

# Part I (1/2) / Method

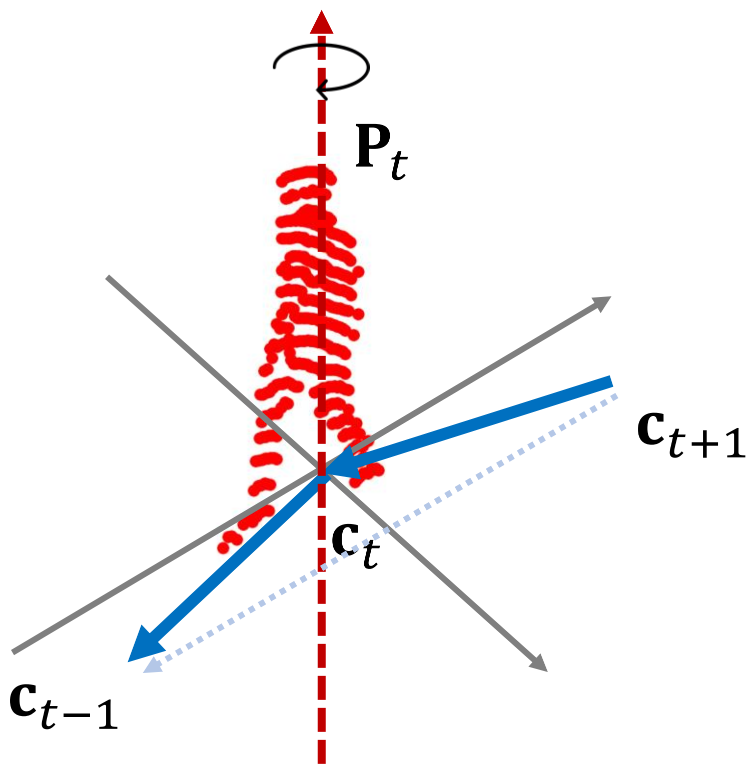
- Outline

| | | |
|---|---|---|
| Input: LiDAR pedestrian sequence data | → Gait direction transformation<br>Input generation<br>Recognition network | → Person Identification |



# Part I (1/2) / Method / Gait Direction Transformation

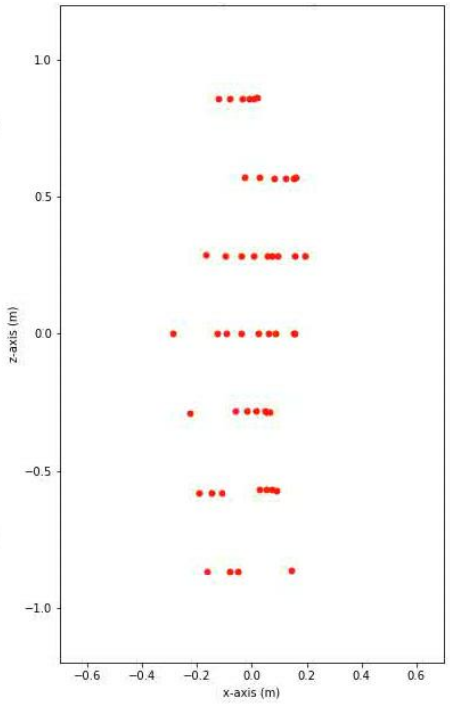
- Obtain a gait directional angle $\theta_t$:
  - $\theta_t = \arctan(c_{t+1,y} - c_{t-1,y}, c_{t+1,x} - c_{t-1,x})$
- Rotate the $\mathbf{P}_t$ around $\mathbf{c}_t$ as the $z$-axis:
  - $\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_t - \pi/2) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t)$
- The case of generating a back-view gait image:
  - $\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_t - \pi) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t)$



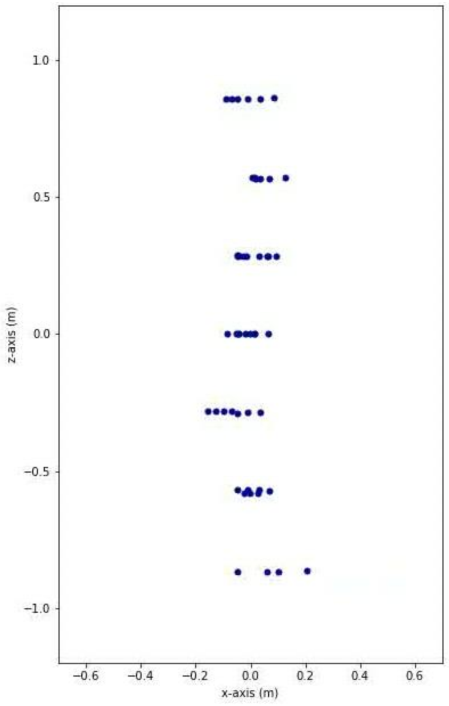

$\mathbf{P}_t$: Original subject point set for the timestep $t$
$\mathbf{c}_t$: Center of gravity for a subject
$\mathbf{R}_z$: Rotation matrix around the $z$-axis
$\hat{\mathbf{P}}_t$: Subject point set transformed

# Part I (1/2) / Method / Gait Direction Transformation

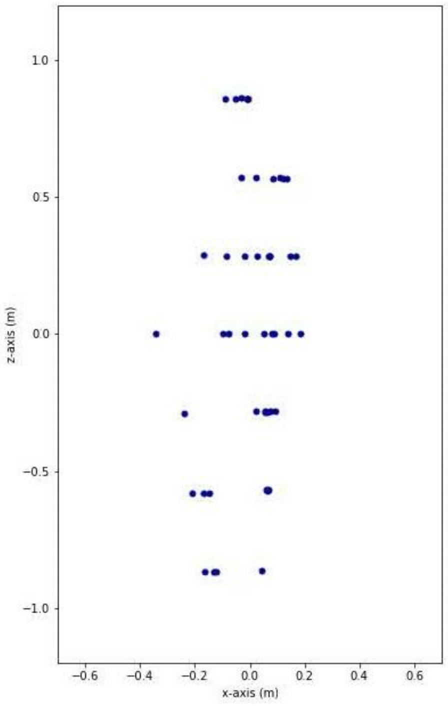
- Examples of GDT



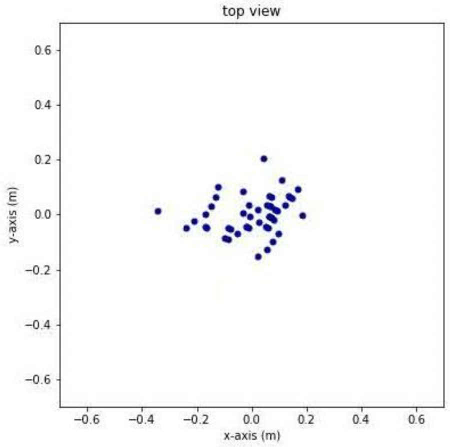

Sensor view




Left-side view



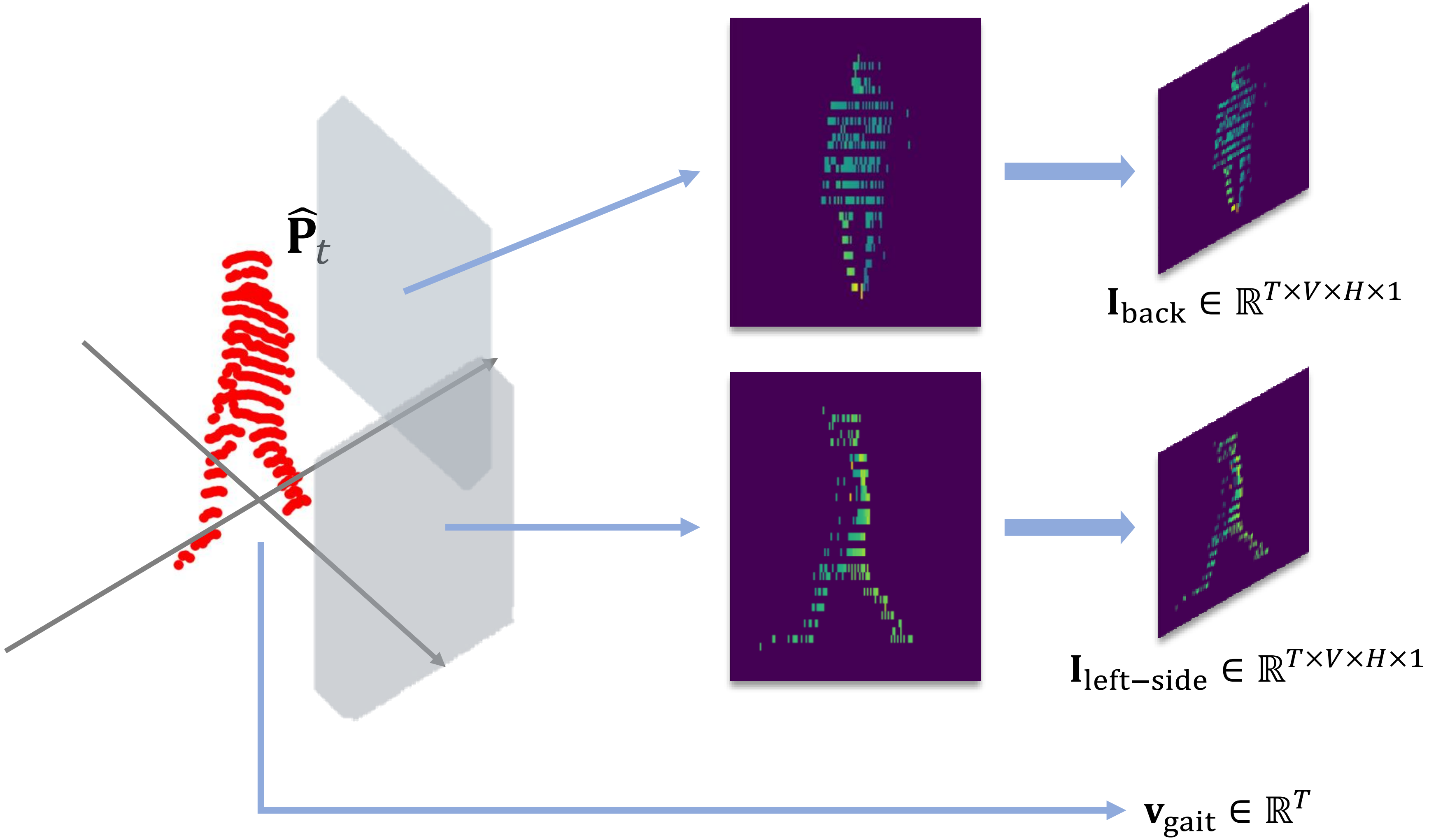

Front view




Bird eyes view

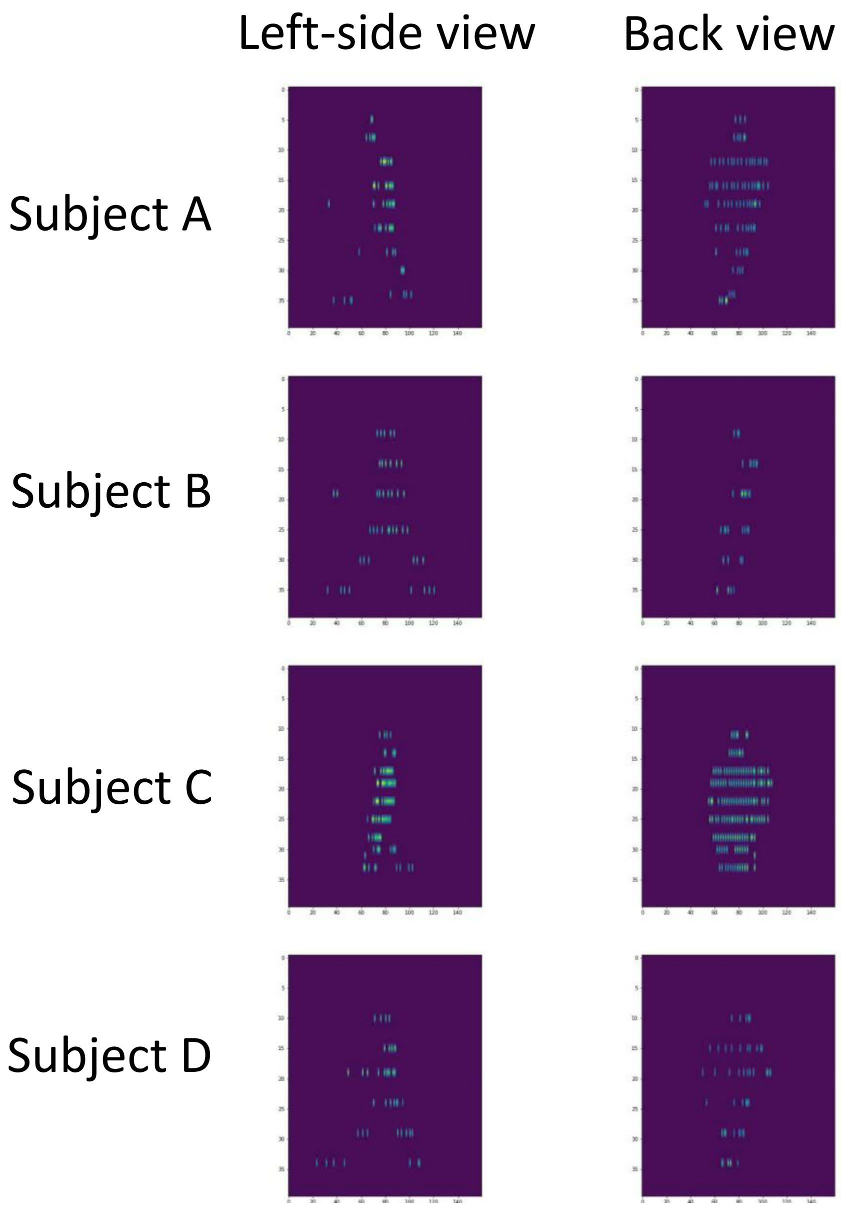
# Part I (1/2) / Method / Input Generation

- Generate three inputs for the proposed network

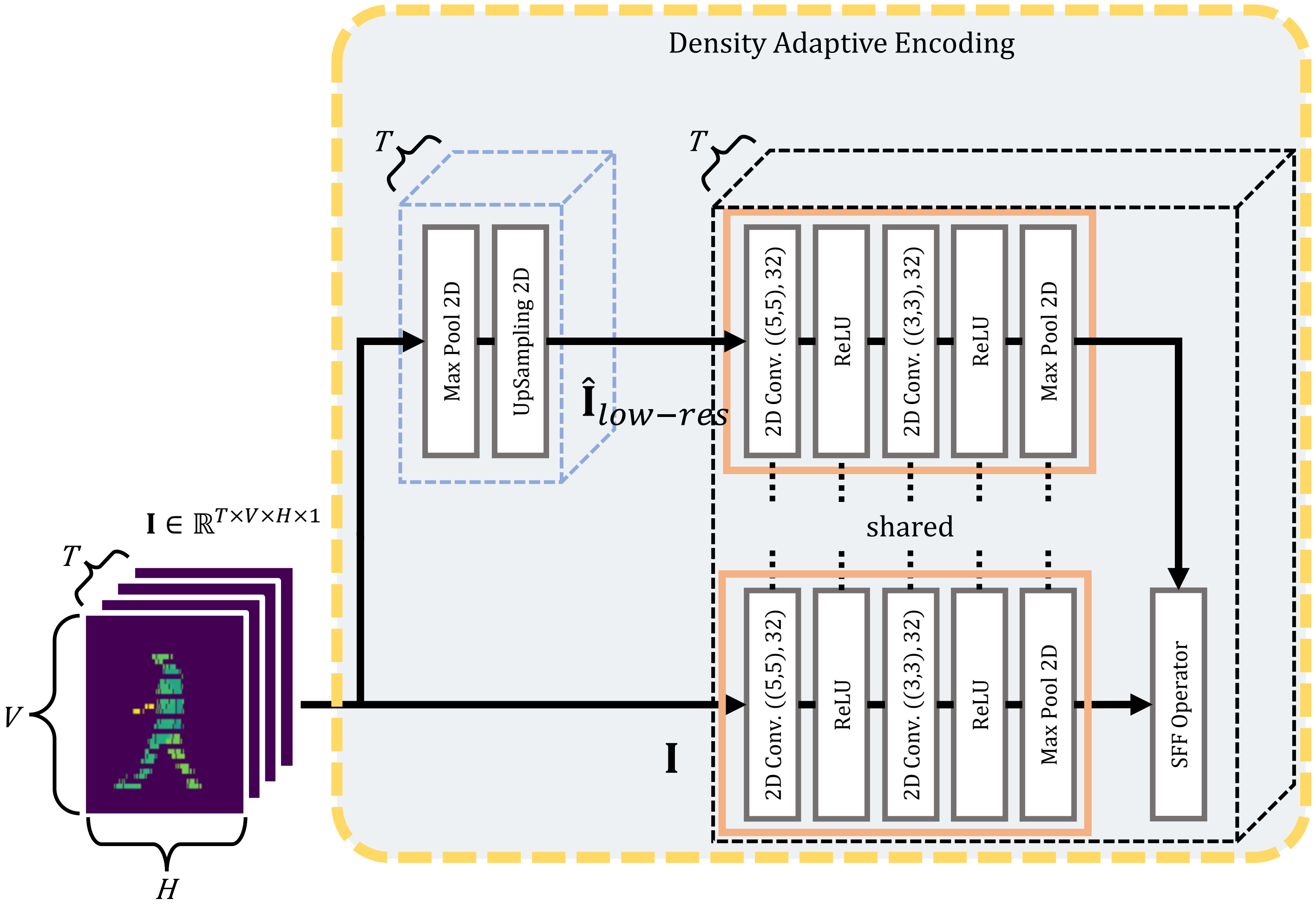


# Part I (1/2) / Method / Input Generation

- Extract gait videos representing the depth information of pedestrians
- Comparing surface depths at each pixel position on the projection plane (Similar to Z-buffer method)
- Obtain the gait image sequence $\mathbf{I} \in \mathbb{R}^{T \times V \times H \times 1}$ and the gait speed sequence $\mathbf{v}_{\text{gait}} \in \mathbb{R}^{T}$



# Part I (1/2) / Method / Network

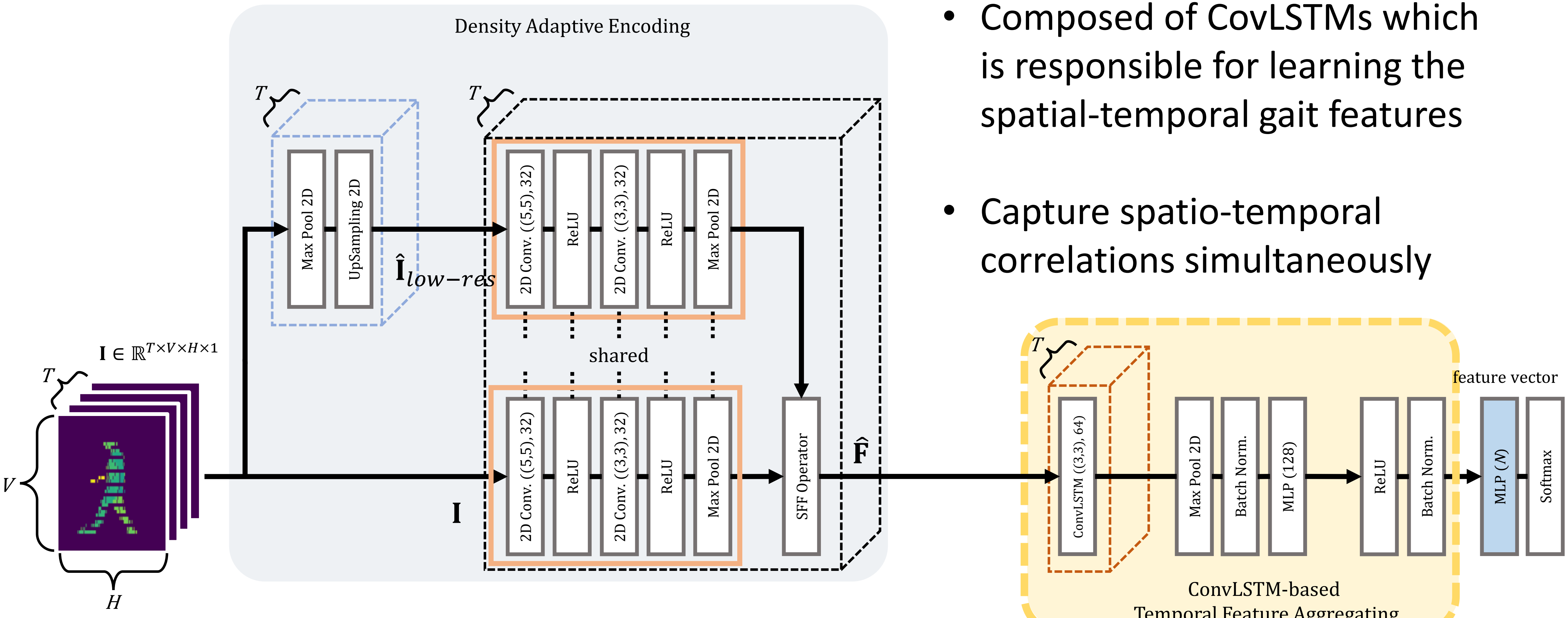



- Leverage the low resolution which may be robust to sparse data and better recognize coarse-grained patterns

- Combine multi-scale features extracted from different resolutions

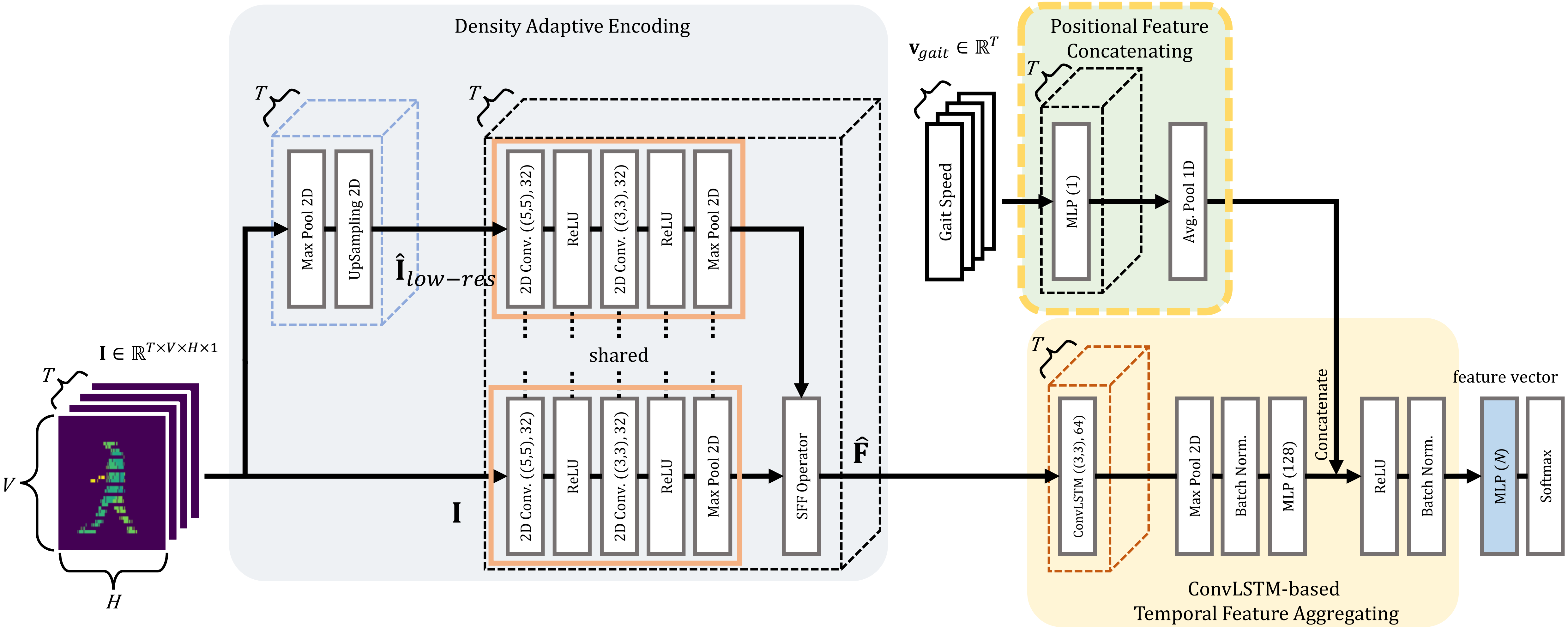
$$\hat{\mathbf{F}} = \frac{1}{2} \cdot (\mathrm{Conv2D}(\mathbf{I}) \oplus \mathrm{Conv2D}(\hat{\mathbf{I}}_{low-res}))$$

# Part I (1/2) / Method / Network




- Composed of CovLSTMs which is responsible for learning the spatial-temporal gait features
- Capture spatio-temporal correlations simultaneously

# Part I (1/2) / Method / Network

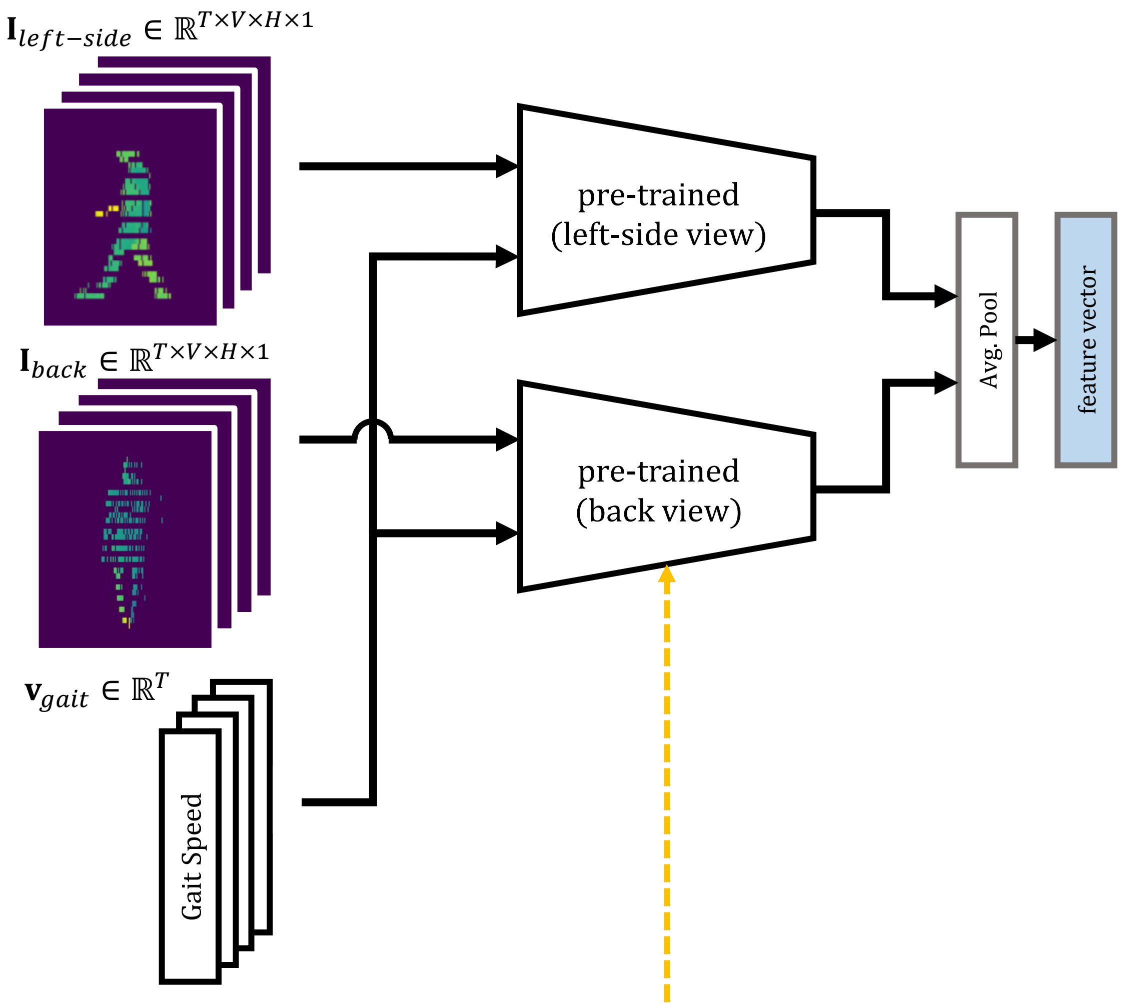


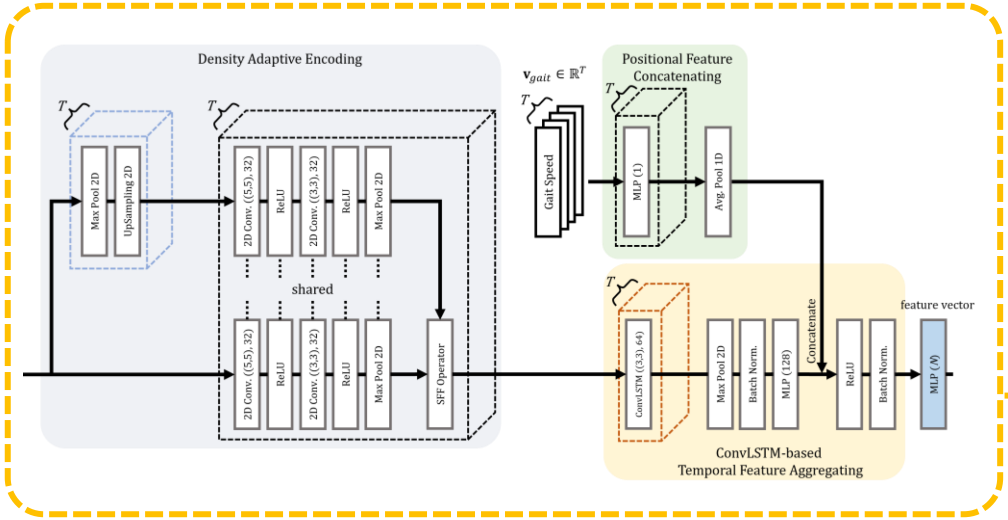

- Take advantage of the walking speed information

# Part I (1/2) / Method / Network

- Obtain more discriminative features from two viewpoints: $\mathrm{I}_{\mathrm{left-side}}$ and $\mathrm{I}_{\mathrm{back}}$

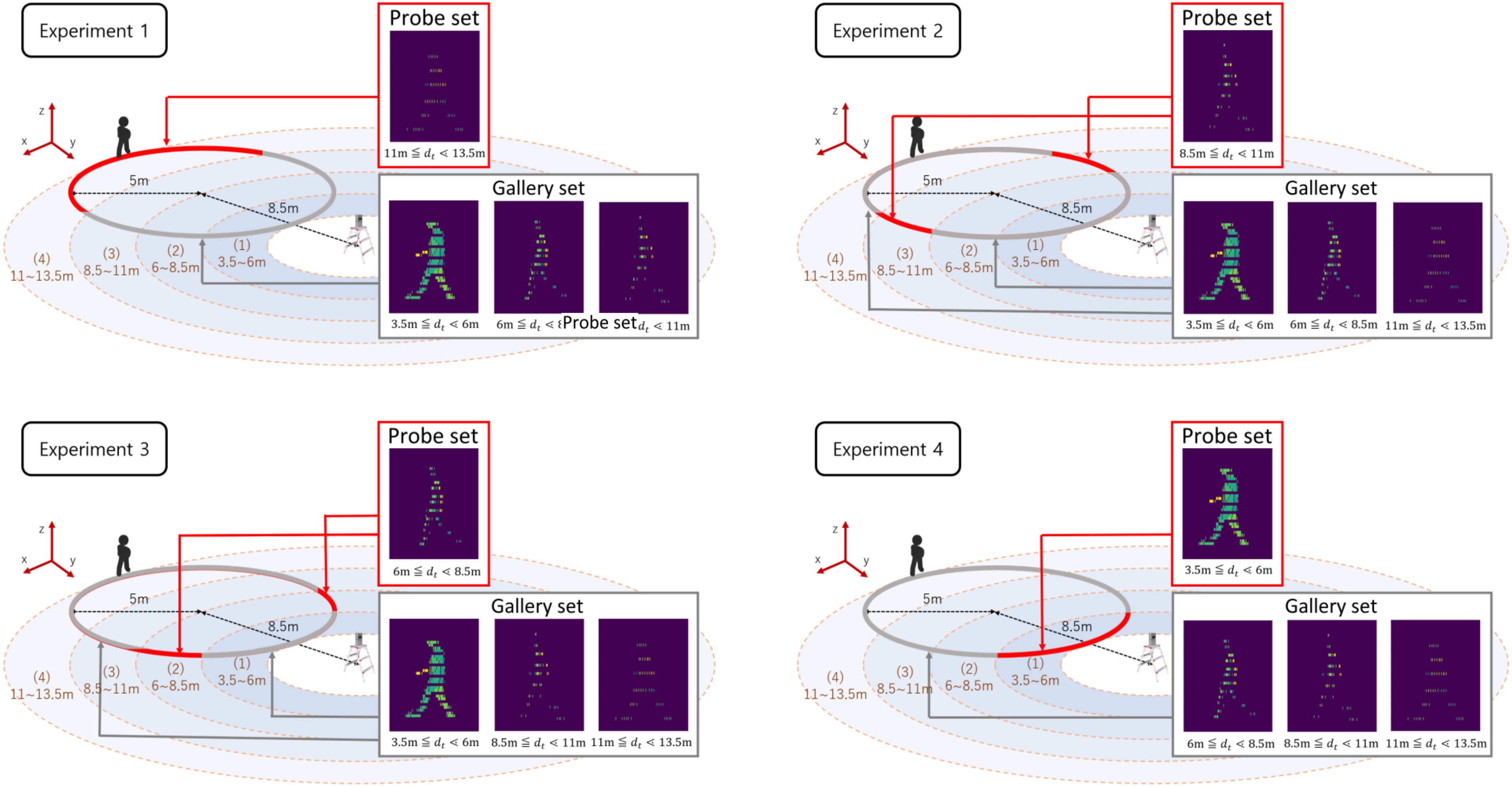
- Aggregate the outputs of two netwo rks pre-trained on different viewpoi nts



# Part I (1/2) / Experiments / Implementation Details

- Four combinations of four subsets were used for testing



# Part I (1/2) / Experiments / Implementation Details

- Each dataset contains 31 subjects
  - Training set : **first 16 subjects**
  - Test set: **remaining 15 subjects**

- Identify subjects using the **Nearest Neighbor Algorithm (Rank-1)**
  - Compute **cosine similarity** between the **gallery** and **probe**

- Settings:
  - Loss function: Cross-entropy loss
  - Optimization: RMSProp
  - Learning rate: 0.001
  - Training batch size: 20
  - Regularization: early stopping
    - patience: 20

| | |
|---|---|
| Total training set | 4 * 140 * 16 = 8,960 |
| Total val. set | 4 * 35 * 16 = 2240 |
| Gallery set (in each pattern) | 3 * 140 * 15 = 6300 |
| Probe set (in each pattern) | 1 * 35 * 15 = 525 |

# Part I (1/2) / Experiments / Ablation Study

TABLE I: Averaged rank-1 accuracies on our dataset. The recognition accuracy in which the range of the test set is not included in range of the training sets is shown in bold.

| Network | Gallery | | | | Probe | | | | mean | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3.5–6m | 6–8.5m | 8.5–11m | 11–13.5m | 3.5–6m | 6–8.5m | 8.5–11m | 11–13.5m | included | non-included |
| 2V-Gait (ours) → TFA | ✓ | ✓ | ✓ | | 89.90 | 91.40 | 88.57 | **62.67** | 87.39 | **72.60** |
| | ✓ | ✓ | | ✓ | 89.33 | 91.59 | **73.52** | 81.71 | | |
| | ✓ | | ✓ | ✓ | 88.76 | **77.44** | 86.10 | 80.57 | | |
| | | ✓ | ✓ | ✓ | **76.76** | 91.01 | 86.48 | 83.24 | | |
| 2V-Gait (ours) → TFA + DAE | ✓ | ✓ | ✓ | | 89.71 | 91.59 | 89.52 | **68.00** | 87.80 | **74.04** |
| | ✓ | ✓ | | ✓ | 89.52 | 89.87 | **71.62** | 82.48 | | |
| | ✓ | | ✓ | ✓ | 88.95 | **85.47** | 87.81 | 81.90 | | |
| | | ✓ | ✓ | ✓ | **71.05** | 91.01 | 87.62 | 83.62 | | |
| 2V-Gait (ours) → TFA + DAE + PFC | ✓ | ✓ | ✓ | | 81.33 | 89.29 | 83.62 | **69.71** | 84.26 | **71.65** |
| | ✓ | ✓ | | ✓ | 82.86 | 89.29 | **66.86** | 83.05 | | |
| | ✓ | | ✓ | ✓ | 81.14 | **77.44** | 83.05 | 82.48 | | |
| | | ✓ | ✓ | ✓ | **72.57** | 86.04 | 82.10 | 84.76 | | |
| 2V-Gait (ours) → TFA + DAE + PFC + VFA | ✓ | ✓ | ✓ | | 92.95 | 95.22 | 94.86 | **76.57** | 93.57 | **84.27** |
| | ✓ | ✓ | | ✓ | 91.81 | 95.41 | **89.71** | 91.43 | | |
| | ✓ | | ✓ | ✓ | 92.38 | **89.29** | 95.81 | 90.67 | | |
| | | ✓ | ✓ | ✓ | **81.52** | 95.22 | 95.62 | 91.43 | | |

- Achieved a better performance by gradually adding the proposed modules

# Part I (1/2) / Experiments / Main Results

| Network | Gallery | | | | Probe | | | | mean | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3.5–6m | 6–8.5m | 8.5–11m | 11–13.5m | 3.5–6m | 6–8.5m | 8.5–11m | 11–13.5m | included | non-included |
| 2V-Gait (ours) | ✓ | ✓ | ✓ | | 92.95 | 95.22 | 94.86 | **76.57** | 93.57 | **84.27** |
| | ✓ | ✓ | | ✓ | 91.81 | 95.41 | **89.71** | 91.43 | | |
| | ✓ | | ✓ | ✓ | 92.38 | **89.29** | 95.81 | 90.67 | | |
| | | ✓ | ✓ | ✓ | **81.52** | 95.22 | 95.62 | 91.43 | | |
| GEINet [8] (Shiraga et al.) | ✓ | ✓ | ✓ | | 87.05 | 88.72 | 85.71 | **64.38** | 84.34 | **73.08** |
| | ✓ | ✓ | | ✓ | 87.81 | 88.53 | **72.76** | 75.43 | | |
| | ✓ | | ✓ | ✓ | 87.43 | **78.59** | 83.24 | 79.81 | | |
| | | ✓ | ✓ | ✓ | **76.57** | 87.19 | 84.95 | 76.19 | | |
| LSTMNet [10] (Yamada et al.) | ✓ | ✓ | ✓ | | 74.48 | 76.29 | 70.67 | **51.43** | 70.53 | **61.78** |
| | ✓ | ✓ | | ✓ | 73.14 | 73.23 | **59.62** | 64.57 | | |
| | ✓ | | ✓ | ✓ | 74.10 | **69.02** | 69.14 | 65.33 | | |
| | | ✓ | ✓ | ✓ | **67.05** | 71.89 | 68.00 | 65.52 | | |

- The left-side view gait video $\mathbf{I}_{\text{left-side}}$ was used in two previous networks
- Present a better performance when all components were applied

# Part I (1/2) / Summary

- The first attempt to develop a LiDAR-based gait recognition model aimed at enhancing robustnesss against variations in distance and walking direction

- Enhance discriminative performance through:
  - Invariant multi-view projection

- Generalize gait features under variations in data sparsity variations through:
  - Multi-scale spatial encoding
  - Walking speed encoding

- Build a LiDAR gait dataset and demonstrate the effectiveness of the proposed identifier

# Part I (2/2): Development of Gait Recognition Models using 3D LiDAR

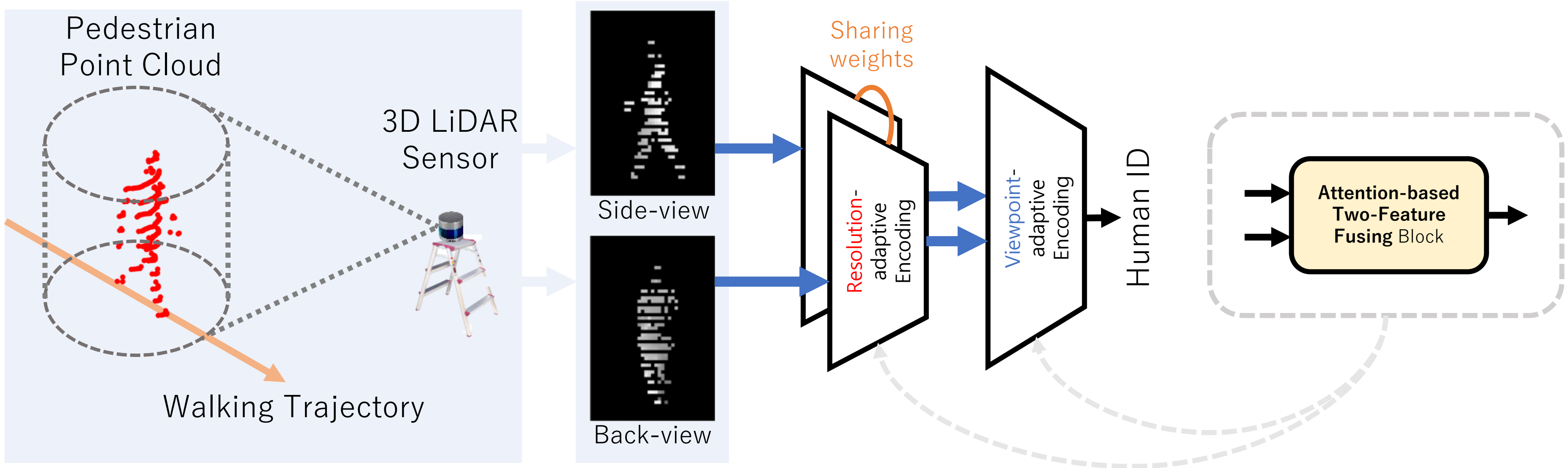
# Part I (2/2) / Motivation

- Challenges in Part I (1/2):
  - **Low inference speed and optimization difficulties** during training
  - Impact of **self-occlusion** on gait shapes
  - The necessary to **independently evaluate the performance** with respect to changes in walking direction and measurement distance/sparsity

- Approaches:
  - Design **a novel attention block** more adaptively to fuse two features for **invariant viewpoint** and **spatial encoding** in an **end-to-end manner**
  - Conduct **an in-depth ablation study** to evaluate the effectiveness of the proposed modules

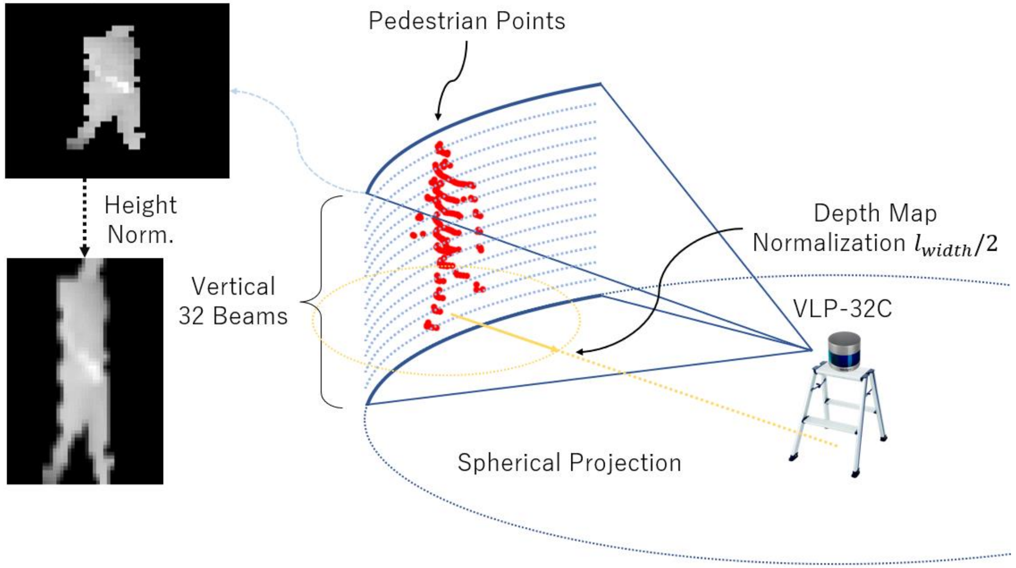
# Part I (2/2) / Method

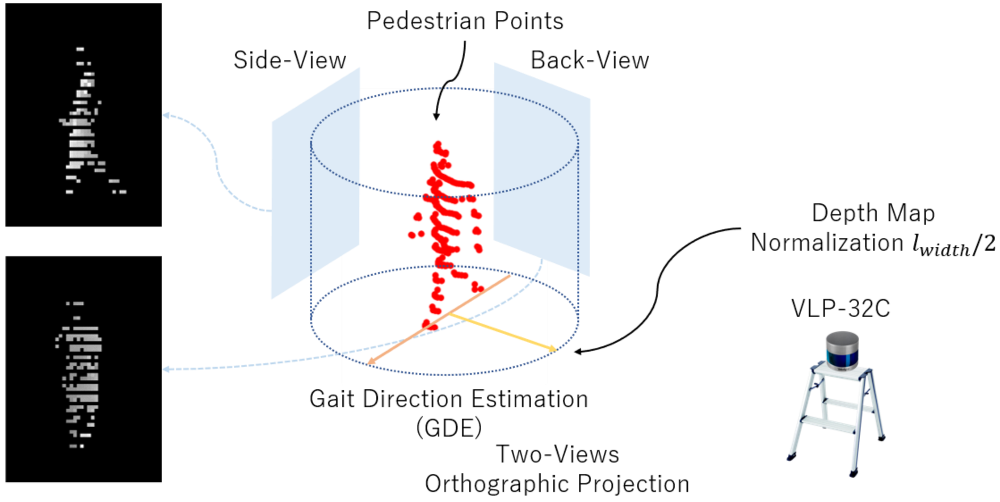
- Overview



# Part I (2/2) / Method / Projection

- LiDAR projection comparison



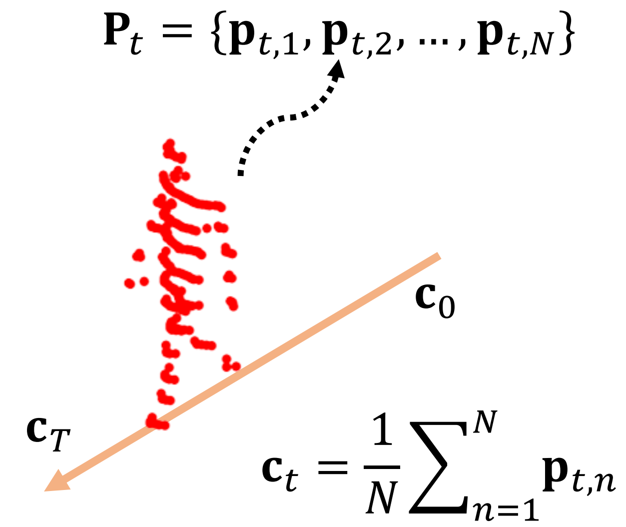



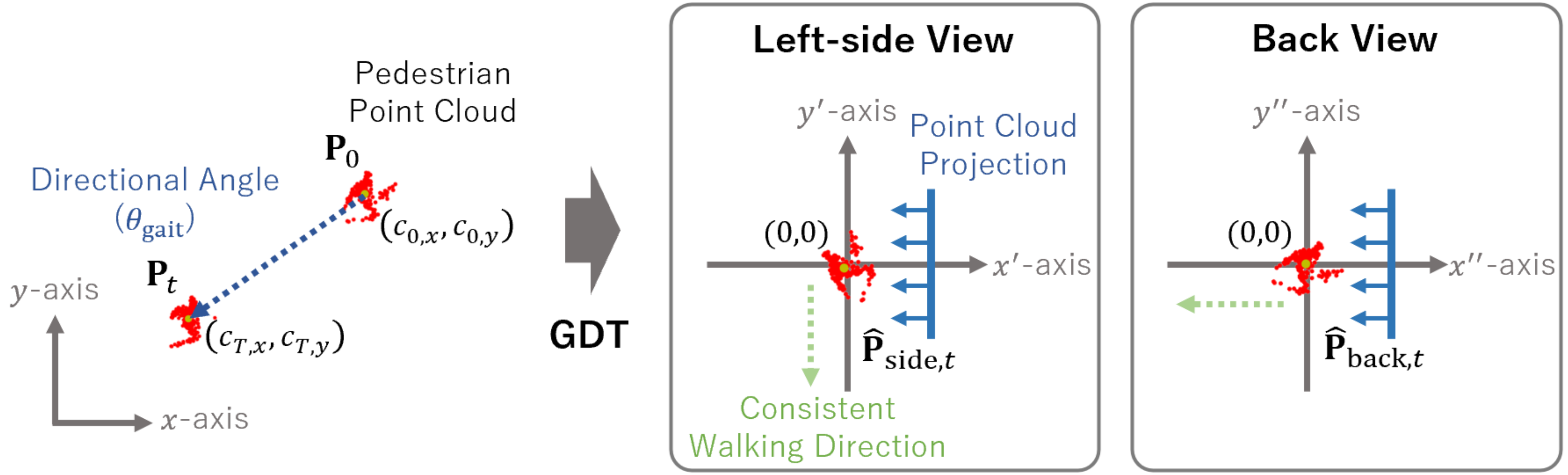

Spherical projection

Orthographic projection (proposed)

# Part I (2/2) / Method / Gait Direction Transformation

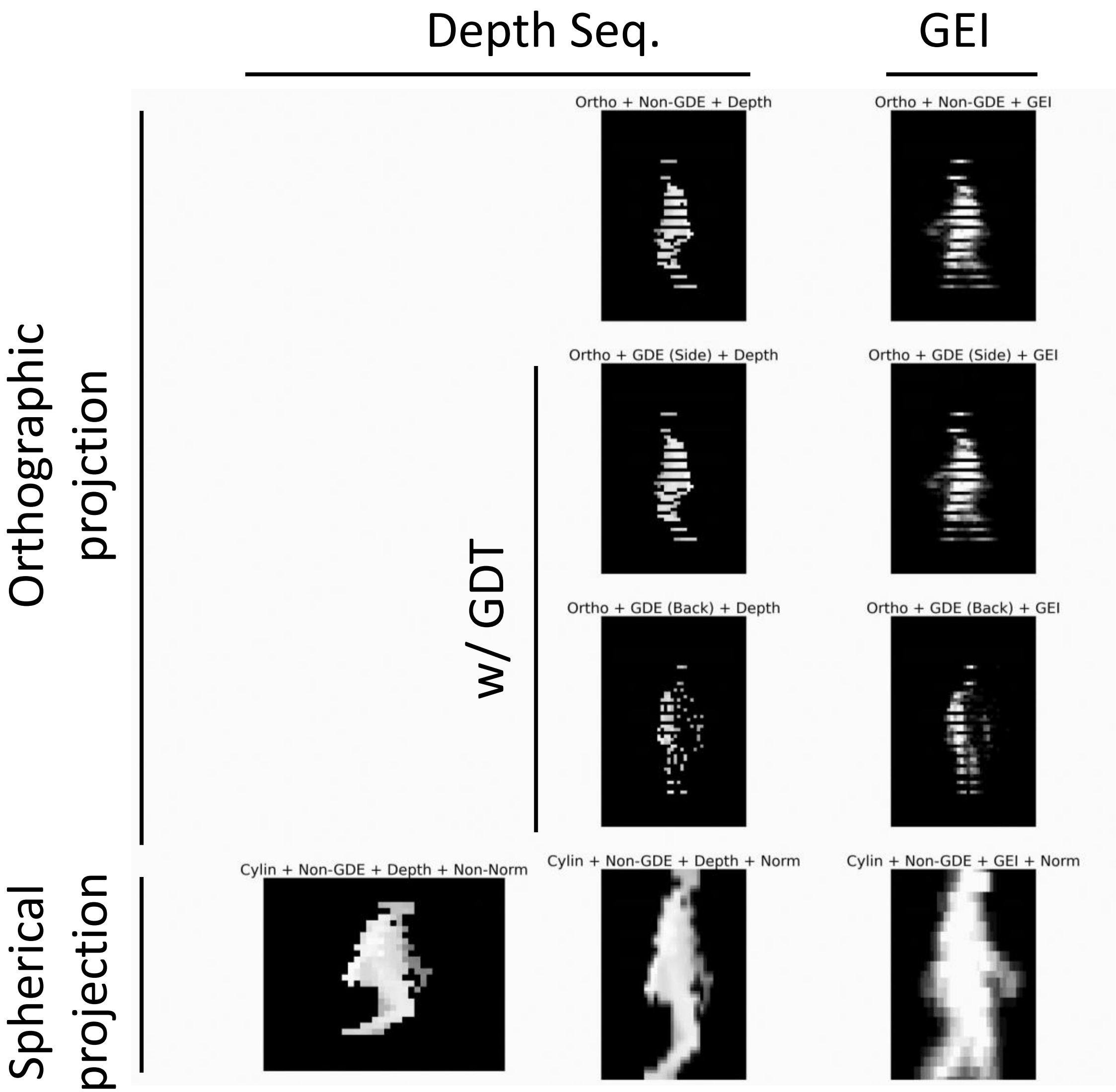
- Obtain a gait directional angle $\theta_{\text{gait}}$:
  - $\mathbf{c}_{\text{gait}} = \mathbf{c}_T - \mathbf{c}_0$
  - $\theta_{\text{gait}} = \arctan(c_{\text{gait},y}, c_{\text{gait},x})$
- Rotate the $\mathbf{p}_{t,n}$ around $\mathbf{c}_t$ as the $z$-axis:
  - $\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_{\text{gait}} - \pi/2) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t)$



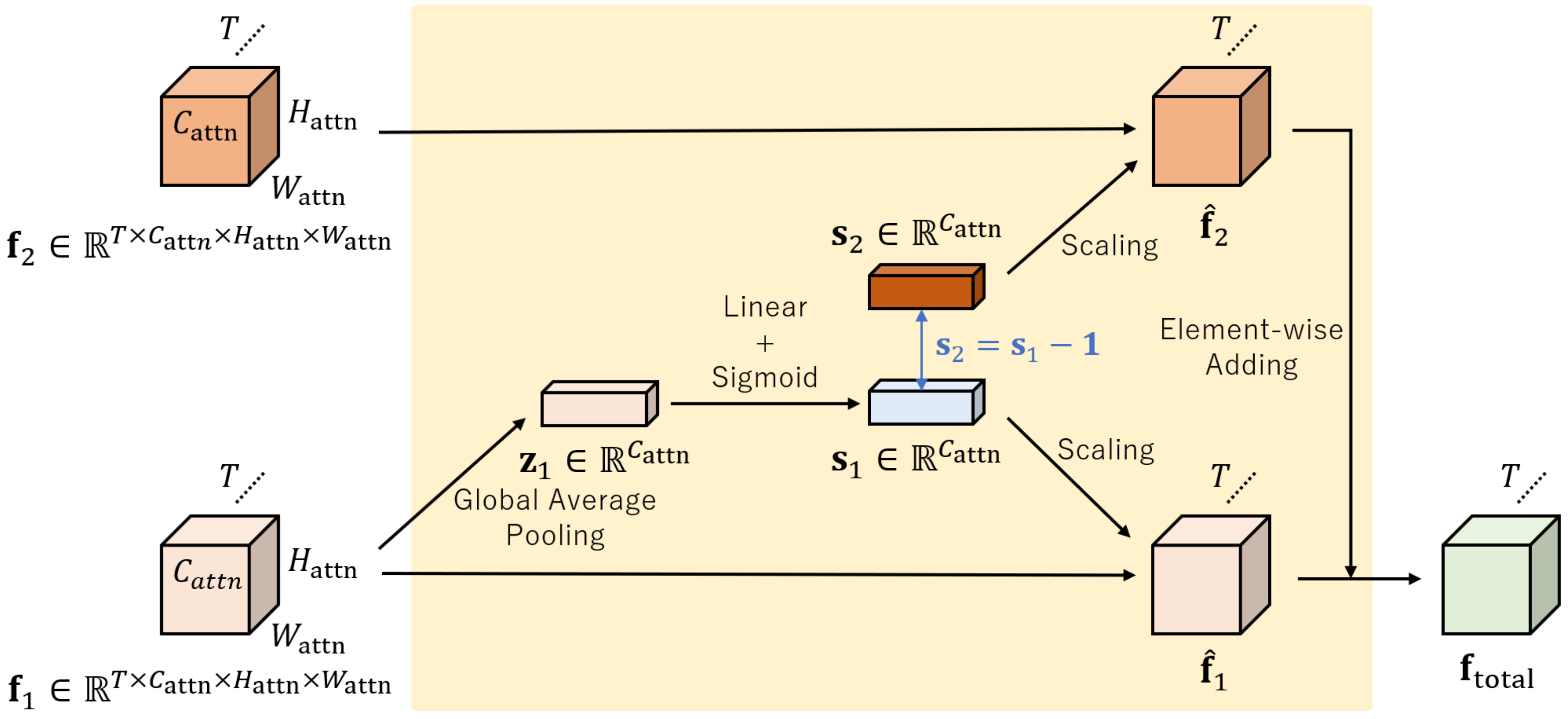
# Part I (2/2) / Method / Gait Direction Transformation

- Examples of GDT



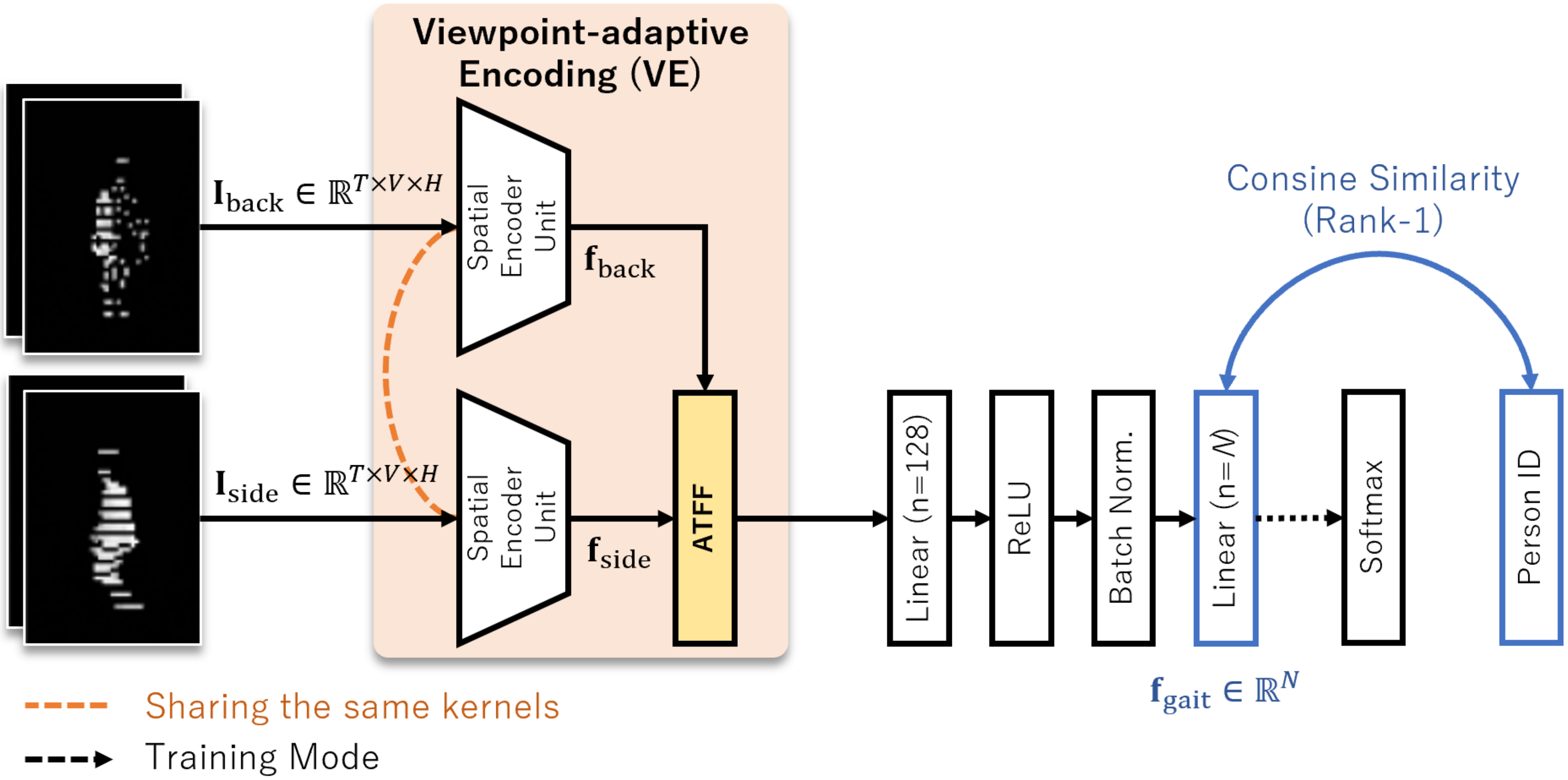
# Part I (2/2) / Attention-based Two-feature Fusing

- Architecture of an **ATFF block**
  - An extension of **SENets**[Hu+, CVPR'18] designed **to fuse two similar feature vectors**



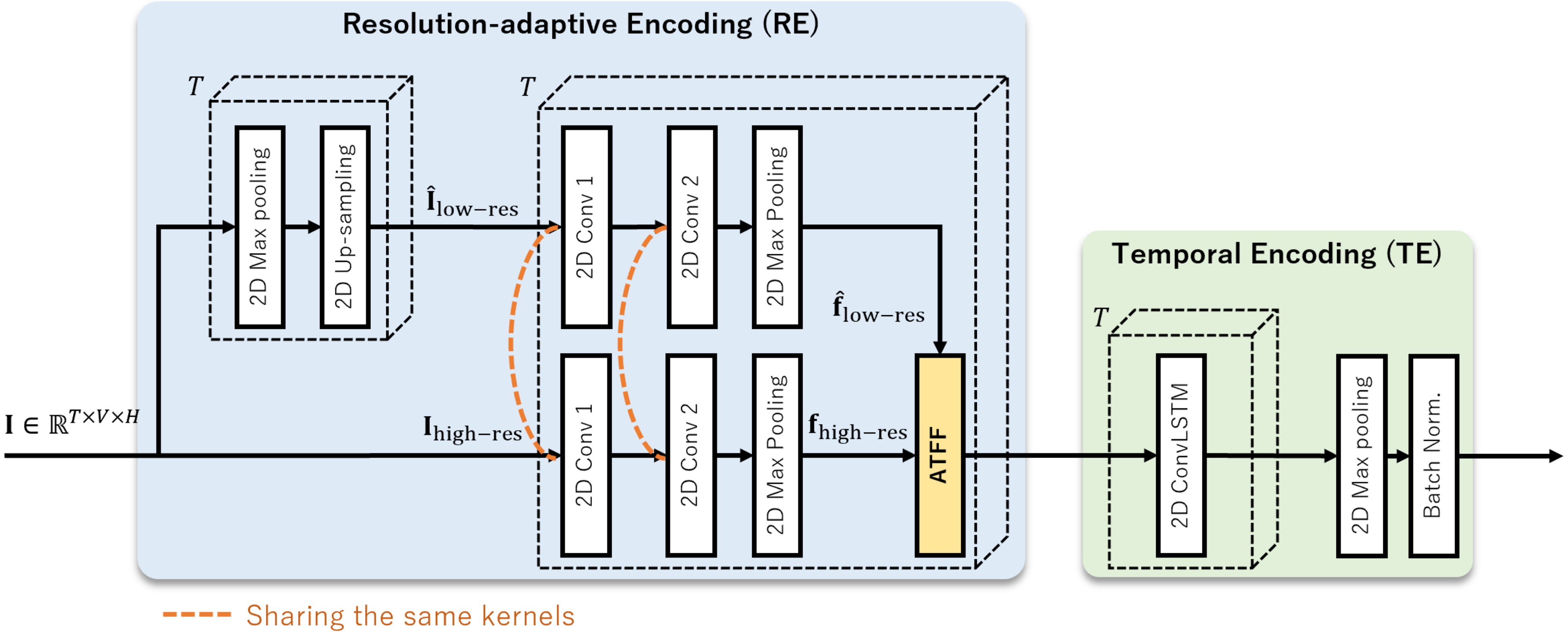
# Part I (2/2) / Method / Network

- Architecture of **overall recognition network**



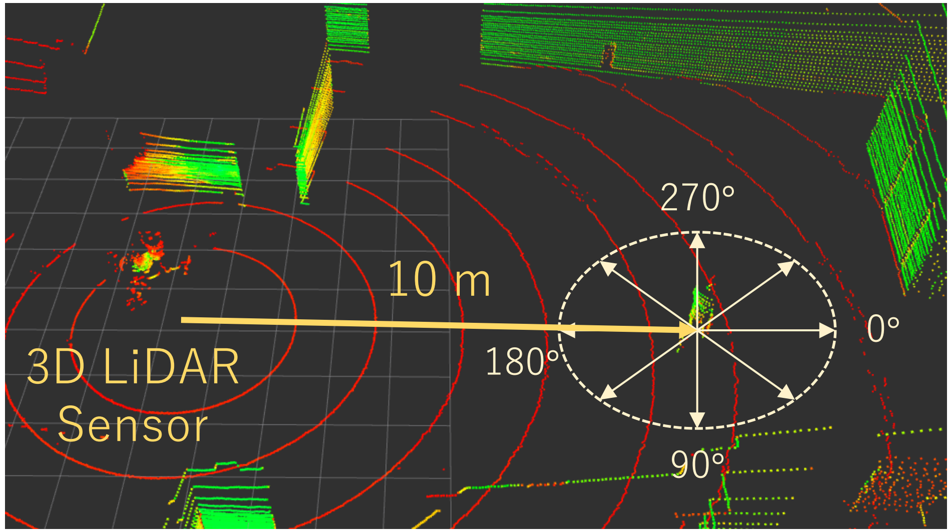
# Part I (2/2) / Method / Network

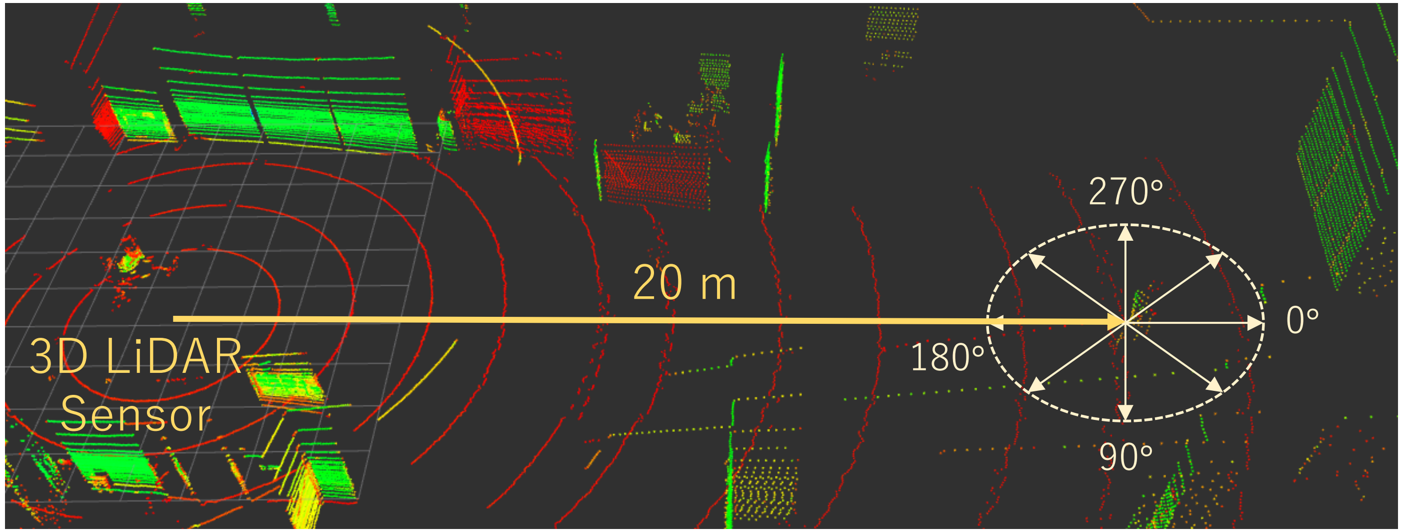
- Architecture of **spatial encoder unit**



# Part I (2/2) / Experiments / Datasets

- Captured using a **Velodyne VLP-32C**
- Gait sequence data collected from **30 subjects**
- Distances: 10 m, 20 m
- Angles: 0°, 45°, 90°, 135°, 180°, 225°, 270°, 315°




Visualization of data acquisition environment

# Part I (2/2) / Experiments / Implementment Details

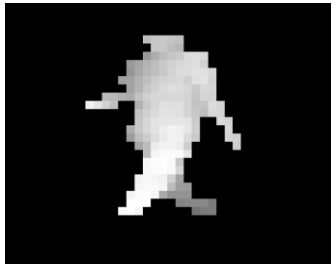
- Each dataset contains 30 subjects
  - Training set : **first 20 subjects**
  - Test set: **remaining 10 subjects**

- Learning settings:
  - Loss function: Cross-entropy loss
  - Optimization: Adam
  - Image size: 64x 44
  - Num. of frames: 15
  - Training batch size: 42
  - Number of training data: 20 * 2 * 8 * 126 = 40,320
  - Iterations: (40,320/42) * 50 = 48,000
  - Height norm. (Spher.): Linear Interpolation

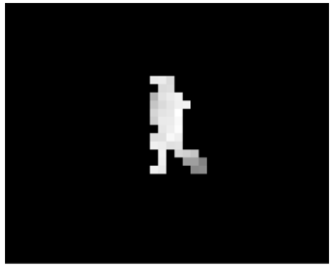
# Part I (2/2) / Experiments / Main Results

- Gallery: 10 m and Probe: 20 m

| Networks | Modalities | Projections | Viewpoints | Means |
|---|---|---|---|---|
| Benedek et al. | GEI | Spher. | Sensor | 30.5 |
| | | | Side | 32.2 |
| | | Ortho. | Sensor | 38.7 |
| | | | Side | 54.9 |
| Shiraga et al. | Depth Seq. | Spher. | Sensor | 30.0 |
| | | | Side | 13.1 |
| | | Ortho. | Sensor | 42.1 |
| | | | Side | 52.3 |
| Proposed | Depth Seq. | Ortho. | Sensor | 69.1 |
| | | | Side | 71.1 |
| | | | Back | 74.8 |
| | | | Side + Back | 81.7 |



Gallery

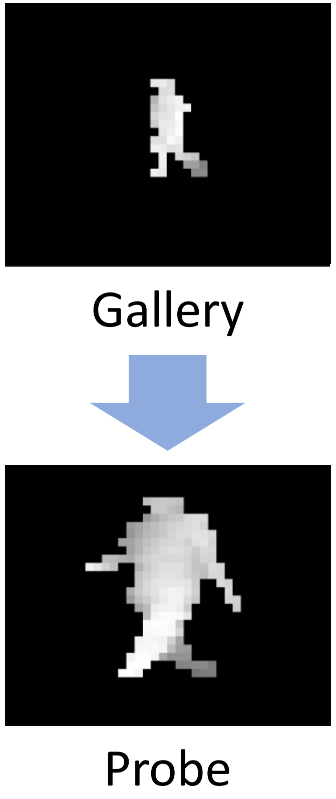


Probe

# Part I (2/2) / Experiments / Main Results

- Gallery: 20 m and Probe: 10 m

| Networks | Modalities | Projections | Viewpoints | Means |
|---|---|---|---|---|
| Benedek et al. | GEI | Spher. | Sensor | 27.5 |
| | | | Side | 43.1 |
| | | Ortho. | Sensor | 53.3 |
| | | | Side | 55.9 |
| Shiraga et al. | Depth Seq. | Spher. | Sensor | 31.0 |
| | | | Side | 36.7 |
| | | Ortho. | Sensor | 38.0 |
| | | | Side | 61.3 |
| Proposed | Depth Seq. | Ortho. | Sensor | 75.2 |
| | | | Side | 75.7 |
| | | | Back | 80.7 |
| | | | Side + Back | 80.8 |



# Part I (2/2) / Experiments / Ablation Study

- Modality and TE

Table 3.4: Effect of input modalities and temporal aggregating manners (%)

| Modalities | | Temporal Encoding (TE) | | Mean |
|---|---|---|---|---|
| Silhouette Seq. | Depth Seq. | 1D-LSTM | ConvLSTM [57] | |
| ✓ | | hidden size = 256 | | 49.2 |
| ✓ | | hidden size = 512 | | 58.4 |
| ✓ | | hidden size = 1024 | | 57.6 |
| ✓ | | | kernel size = $3 \times 3$ | 69.7 |
| ✓ | | | kernel size = $5 \times 5$ | 67.1 |
| ✓ | | | kernel size = $7 \times 7$ | 66.2 |
| | ✓ | hidden size = 256 | | 51.8 |
| | ✓ | hidden size = 512 | | 65.2 |
| | ✓ | hidden size = 1024 | | 65.9 |
| | ✓ | | kernel size = $3 \times 3$ | **72.1** |
| | ✓ | | kernel size = $5 \times 5$ | 70.4 |
| | ✓ | | kernel size = $7 \times 7$ | 68.5 |

# Part I (2/2) / Experiments / Ablation Study

- Impact of RE

Table 3.5: Ablation experiment for resolution-adaptive encoding (RE) (%)

| Original Res.($\mathbf{I}_{\text{high-res}}$) | Low Res.($\hat{\mathbf{I}}_{\text{row-res}}$) | Fusion | | | Mean |
|---|---|---|---|---|---|
| | | Methods | $T$-pooling | Attention Targets ($\mathbf{f}_1$) | |
| ✓ | | | | | 63.3 |
| | ✓ | | | | 51.4 |
| ✓ | ✓ | Element-wise Add. | | | 69.9 |
| ✓ | ✓ | Channel-wise Concat. | | | 69.5 |
| ✓ | ✓ | SE-Net [22] | | | 71.4 |
| ✓ | ✓ | ATFF | | Low Res. ($\hat{\mathbf{f}}_{\text{low-res}}$) | 68.7 |
| ✓ | ✓ | ATFF | ✓ | Low Res. ($\hat{\mathbf{f}}_{\text{low-res}}$) | **72.1** |
| ✓ | ✓ | ATFF | ✓ | Original Res. ($\mathbf{f}_{\text{high-res}}$) | 71.8 |

# Part I (2/2) / Experiments / Ablation Study

- Impact of VE

Table 3.6: Ablation experiment for viewpoint-adaptive encoding (VE) (%)

| Original view | Side-view | Back-view | Fusion | Mean |
|---|---|---|---|---|
| ✓ | | | | 72.1 |
| | ✓ | | | 73.4 |
| | | ✓ | | 77.3 |
| | ✓ | ✓ | Average Pooling [1] | 79.1 |
| | ✓ | ✓ | Max Pooling | 78.5 |
| | ✓ | ✓ | Concatenating | 77.3 |
| | ✓ | ✓ | ATTF ($T = 1$) | **81.2** |

# Part I (2/2) / Experiments / Practicality

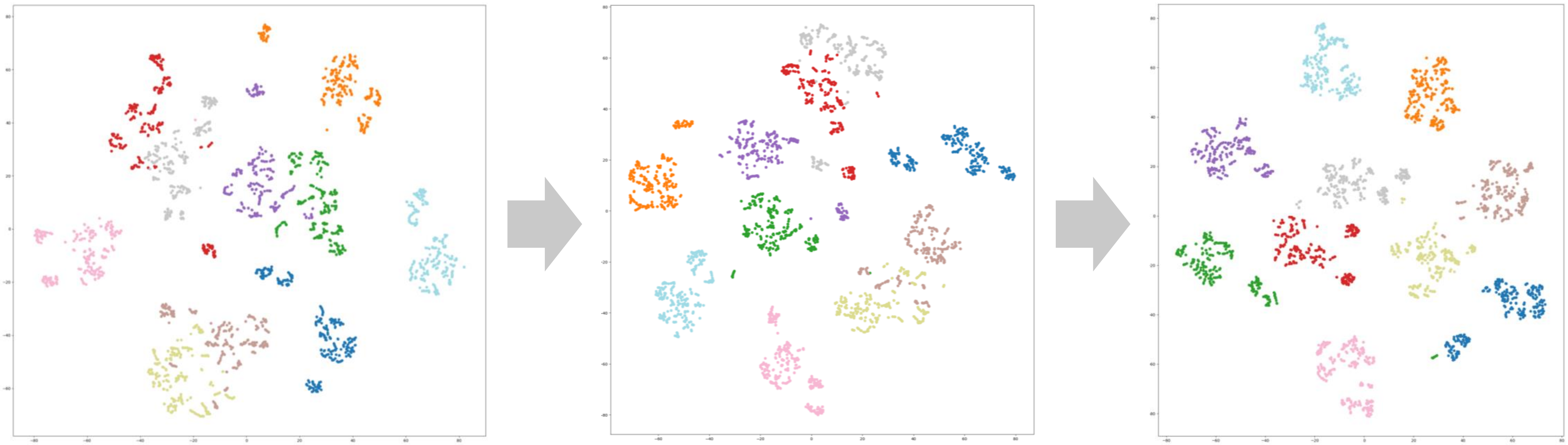
- Quantitative results

Table 3.7: Comparison with prior studies for evaluating practicality by limiting viewing angles (%)

| Networks | Modalities | Projection | Viewpoints | | | Gallery | | |
|---|---|---|---|---|---|---|---|---|
| | | | Sensor-view | Side-view | Back-view | 270 ° (Side-view) | 0 ° (Back-view) | 315 ° (Oblique-view) |
| Benedek et al. [6] | GEI | Spher. | ✓ | | | 26.3 | 36.8 | 25.4 |
| | | | | ✓ | | 38.3 | 37.6 | 40.2 |
| | | Ortho. | ✓ | | | 44.2 | 48.1 | 46.5 |
| | | | | ✓ | | 43.7 | 51.1 | 47.4 |
| Shiraga et al. [59] | GEI | Spher. | ✓ | | | 26.4 | 28.1 | 25.2 |
| | | | | ✓ | | 17.8 | 18.8 | 18.9 |
| | | Ortho. | ✓ | | | 46.5 | 54.3 | 51.5 |
| | | | | ✓ | | 51.2 | 44.7 | 53.3 |
| Yamada et al. (Network 1) [76] | Depth Seq. | Spher. | ✓ | | | 31.0 | 25.3 | 32.3 |
| | | | | ✓ | | 14.4 | 16.2 | 18.0 |
| | | Ortho. | ✓ | | | 53.9 | 48.6 | 50.5 |
| | | | | ✓ | | 33.7 | 45.1 | 45.6 |
| Yamada et al. (Network 2) [76] | Depth Seq. | Spher. | ✓ | | | 31.0 | 28.2 | 33.6 |
| | | | | ✓ | | 15.2 | 15.8 | 17.3 |
| | | Ortho. | ✓ | | | 33.5 | 41.9 | 45.8 |
| | | | | ✓ | | 43.4 | 46.6 | 43.4 |
| Ours | Depth Seq. | Spher. | ✓ | | | 39.1 | 53.4 | 39.5 |
| | | | | ✓ | | 50.8 | 47.5 | 48.3 |
| | | | | | ✓ | 40.4 | 49.6 | 47.0 |
| | | | | ✓ | ✓ | 50.9 | 49.5 | 52.1 |
| | | Ortho. | ✓ | | | 64.3 | 62.4 | 68.9 |
| | | | | ✓ | | 67.8 | 61.3 | 66.6 |
| | | | | | ✓ | 63.3 | 67.7 | 67.4 |
| | | | | ✓ | ✓ | **73.0** | **70.2** | **72.7** |

# Part I (2/2) / Experiments / Feature Visualization

- Visuailze gait features through a 2D manifold space by using t-SNE

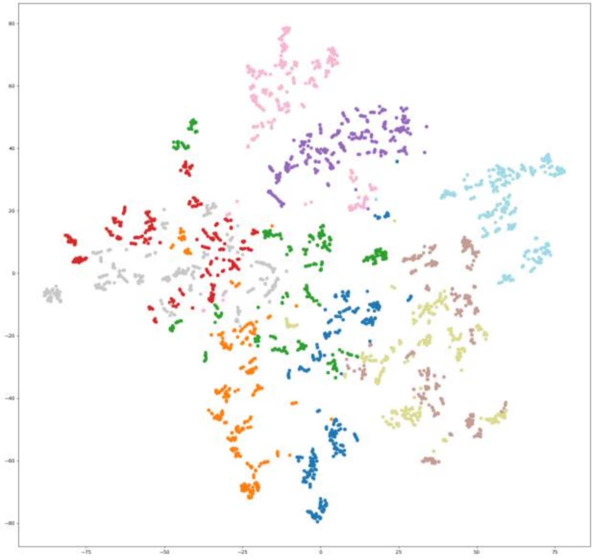


(1) (2) (3)

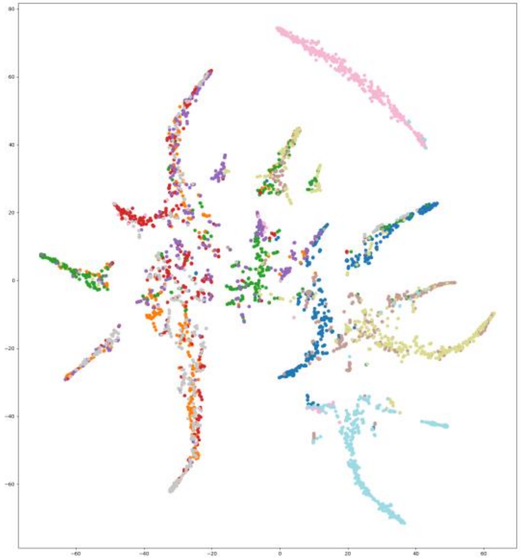
| | RE | VE | Viewpoints |
|---|---|---|---|
| (1) | | | Sensor-view |
| (2) | ✓ | | Sensor-view |
| (3) | ✓ | ✓ | Side- and back-views |

# Part I (2/2) / Experiments / Feature Visualization

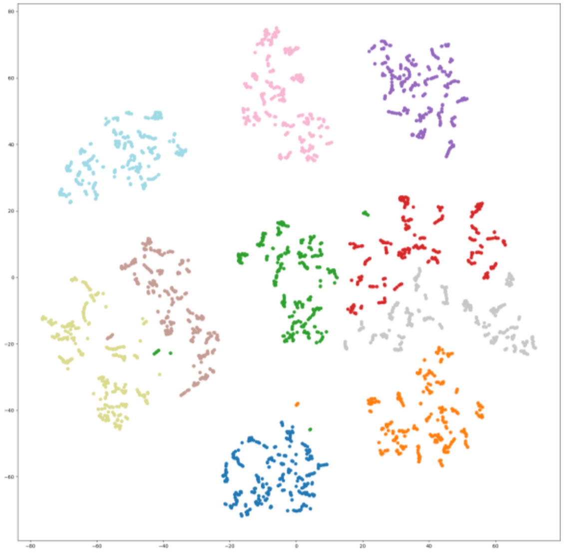
- Feature visualization comparison



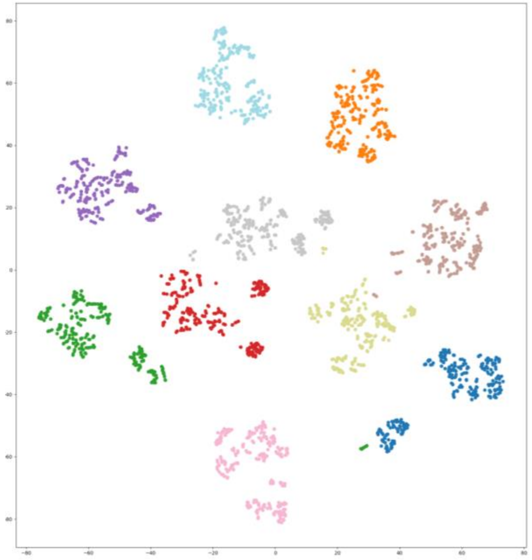
Benedek+, TCVST’20



Yamada+, AR’20



Part I (1/2)



Proposed

# Part I (2/2) / Summary

- Proposed a **attention block** to adaptively fuse two gait features
- Explored in-depth from **three-perspectives**:
  - Point cloud projection
  - Gait direction transformation
  - Recognition network
- Build a LiDAR gait dataset and achieved superior performance of proposed model in both **cross-view** and **cross-distance** condtions

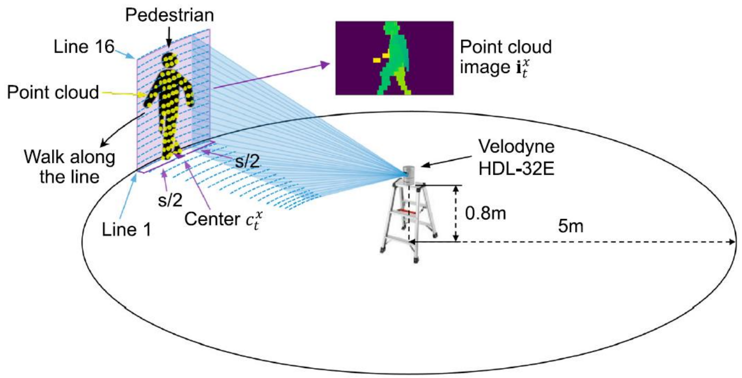
# Part II: Development of Gait Upsampling Models for 3D LiDAR

# Part II / Motivation

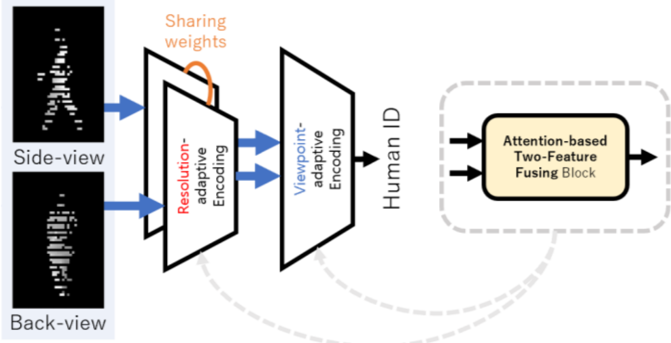
- Recent studies on gait recognition using 3D LiDAR have emerged

Afilliated Lab.

Yamada+



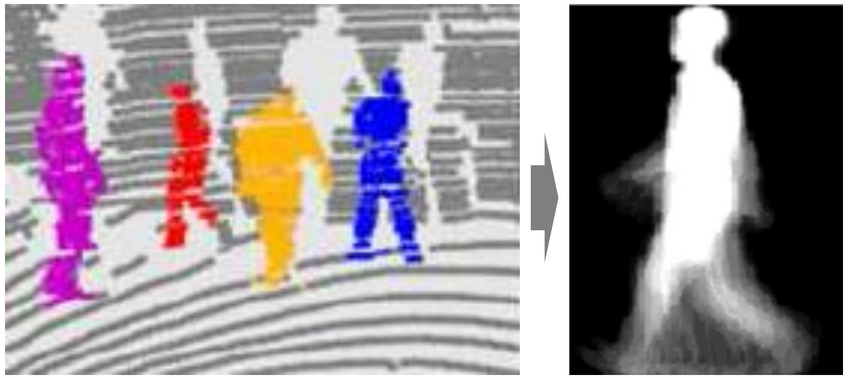

Ahn+



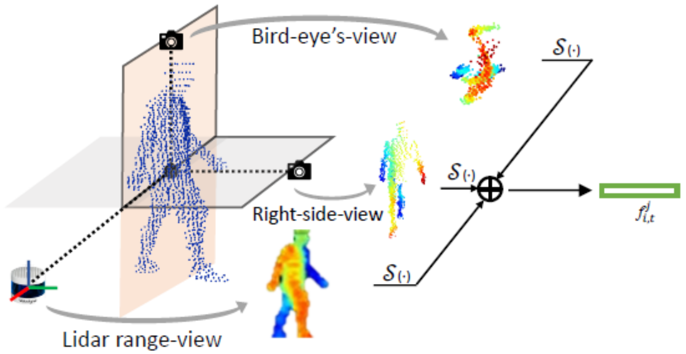

2015 2020 2022 2023 2024 Year

Other Lab.

Benedek+

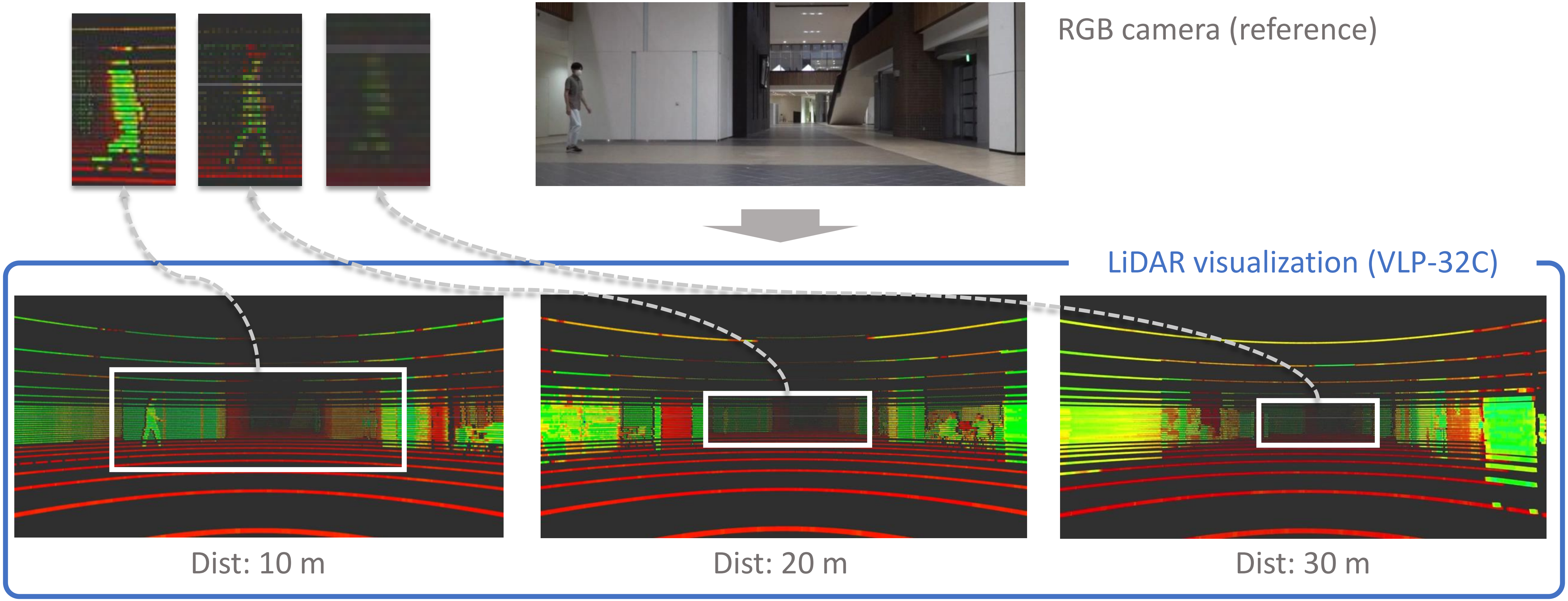


Shen+



# Part II / Motivation

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



# Part II / Motivation

- Changes in **resolution/sparsity** based on **emission patterns (hardware specifications)**

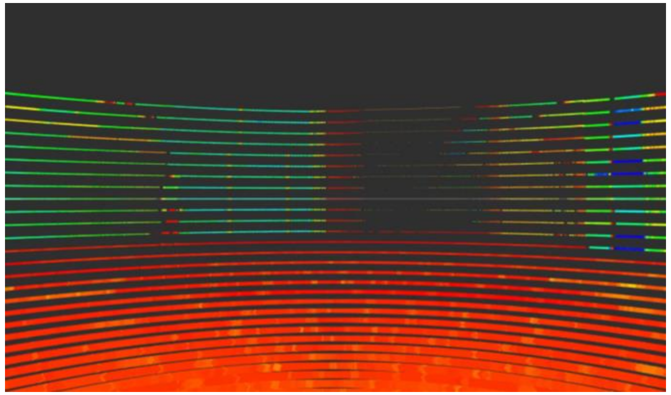
### Mechanical type

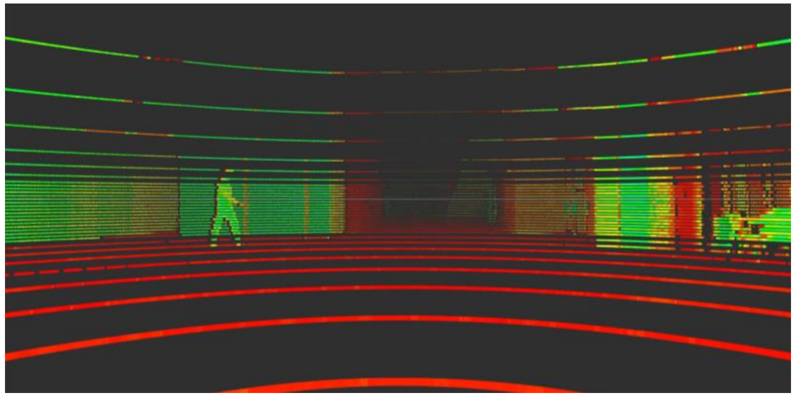
Velodyne HDL-32E


Use 32 identical beams

Velodyne VLP-32C

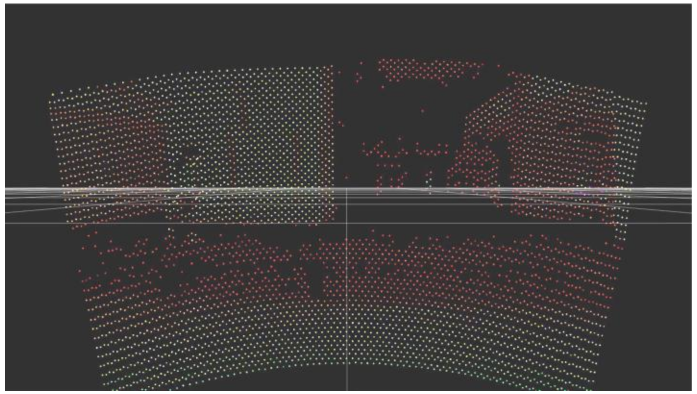

Dist: 10 m





### Solid-state type

Pioneer SSL-S01

# Part II / Motivation

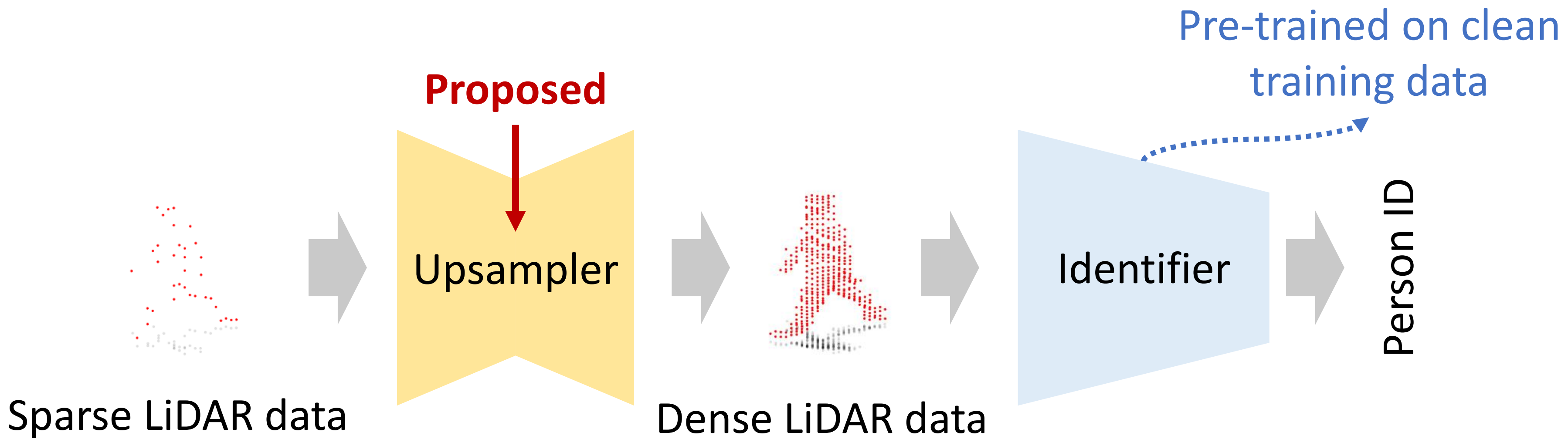
- Challenges:
  - Sparsity of LiDAR data is heavily influenced by **measurement distances** and **hardware specifications**
  - Collecting datasets for all **distances** and **sensor types** is practically difficult

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

# Part II / Motivation

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

- Approaches:
  - Employ **a conditional diffusion model**
  - Restore missing parts of the gait data **through an inpainting strategy**



# Part II / Related Work

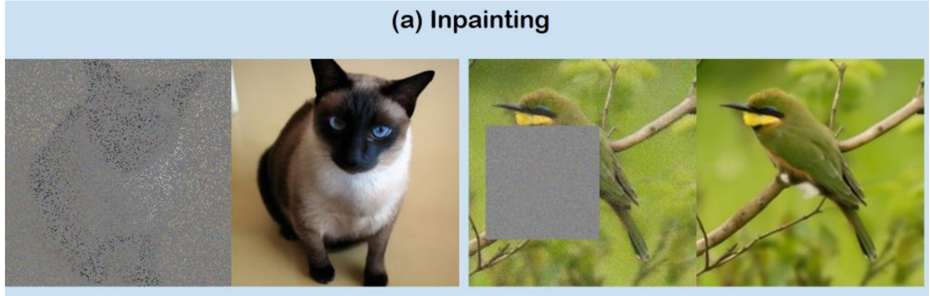
- Typical signal/image restoration (**inpainting**) using **diffusion models**:

Training Phase

**Task-agnostic approach**

GT data → Restoration model

Examples




DPS [Chung+, ICLR’23]

- Learns the underlying distribution and samples data by approximating the posterior
- Tends to be worse than the task-specific approach

**Task-specific approach**

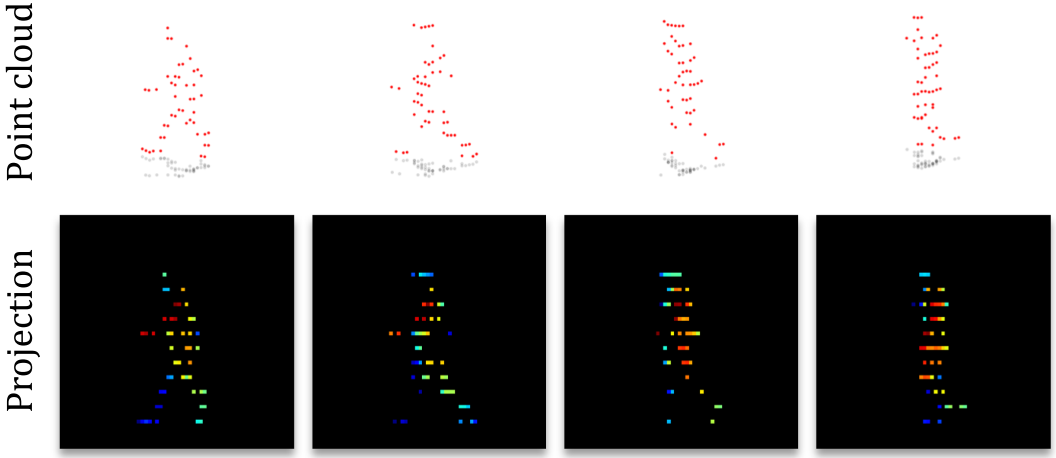
GT data + Noise pattern → Restoration Model


Palette [Saharia+, CVPR’22]

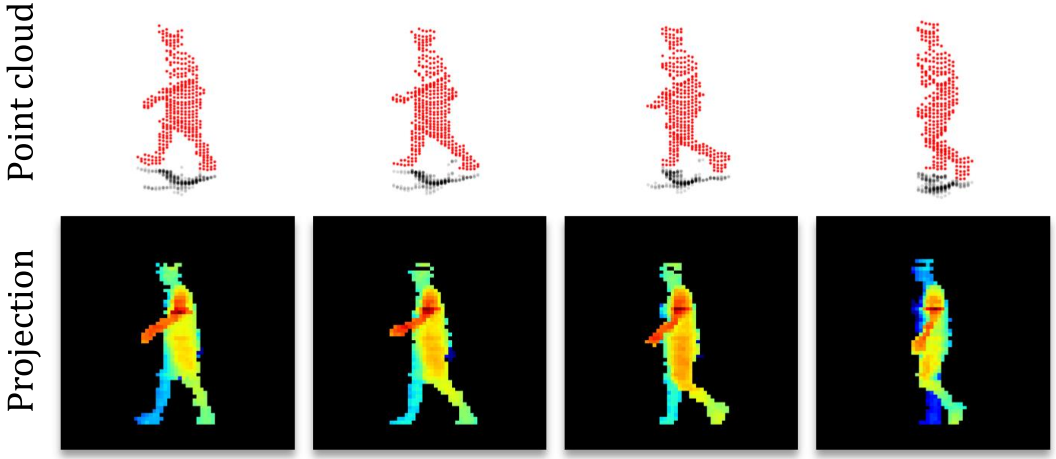
- Conditional diffusion strategy
- Achieves superior performance across various multi-tasks

# Part II / Method

- Overview



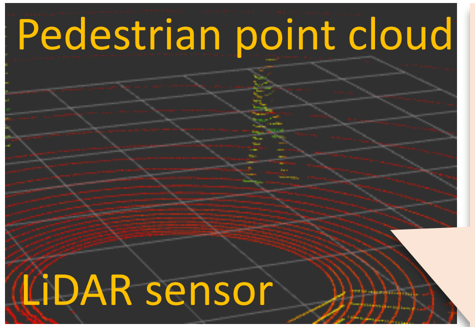

Input: sparse LiDAR gait sequences




Our results

# Part II / Method / Problem Statement

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



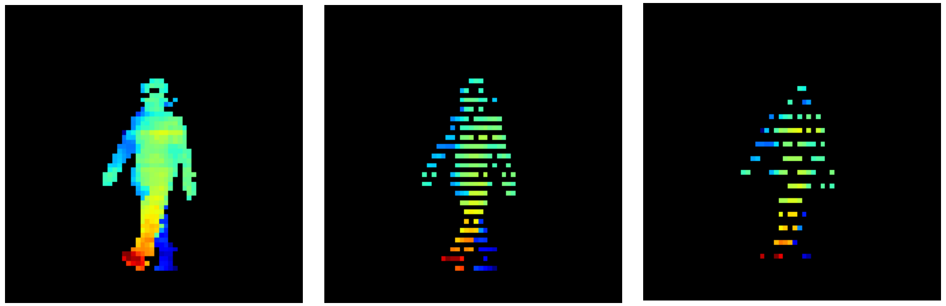

**3D point cloud data** captured by a single LiDAR sensor **cannot** be addressed as **GT data** due to its **self-occlusion**

Degradation noise mask

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

Complete gait video

Orthograhic projection



Captured distance

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

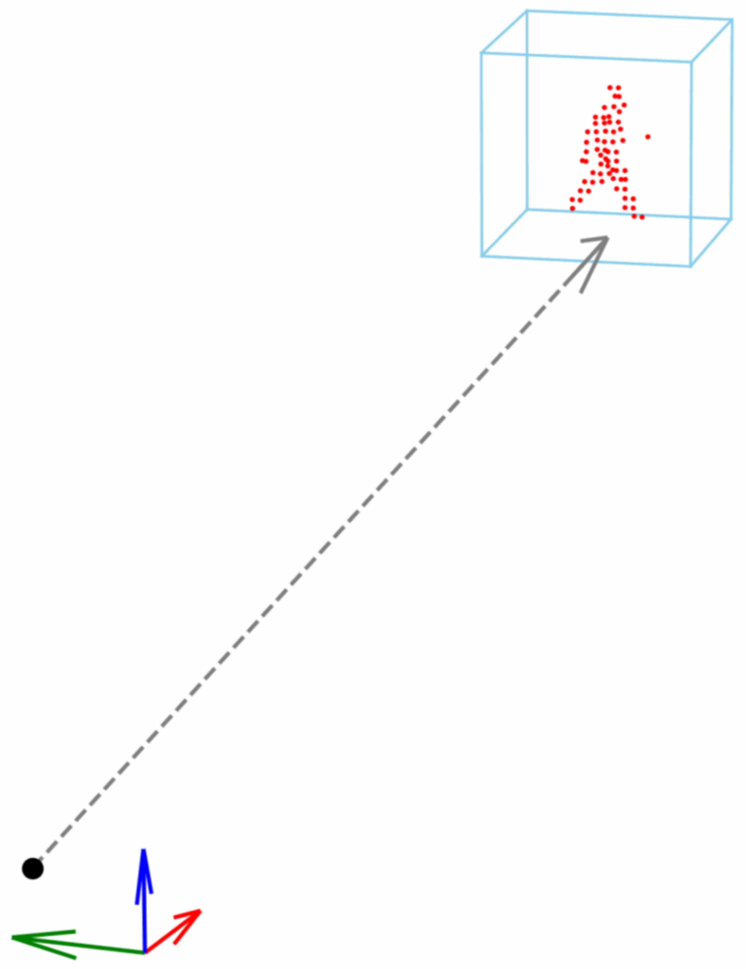
$\mathbf{y}$ $H$ $\mathbf{x}_0$

# Part II / Method / Projection

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

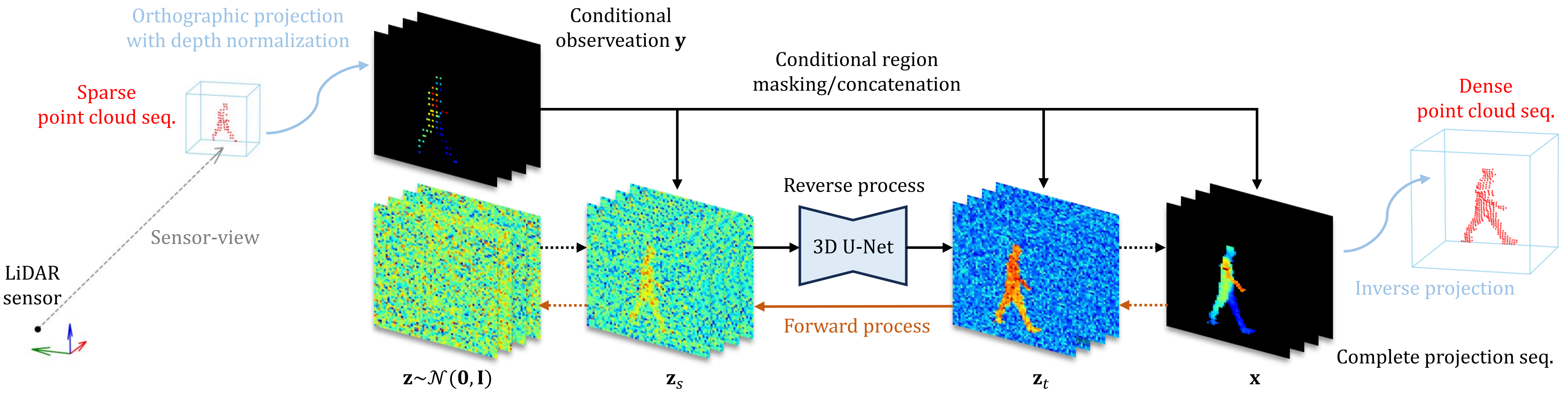
- Obtain the rotated point cloud sequence $\widehat{\mathbf{P}} \in \mathbb{R}^{F\times N\times C}$
  with a directional angle $\theta_{\text{sensor},f}$:
  - $\theta_{\text{sensor},f} = \arctan(c_{f,y}, c_{f,x})$
  - $\widehat{\mathbf{p}}_{f,n} = \left(\mathbf{P}_{f,n} - \mathbf{c}_f\right) \cdot \mathbf{R}_z(\theta_{\text{sensor},f}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

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



# Part II / Method / Network

- Overall of the upsampling network

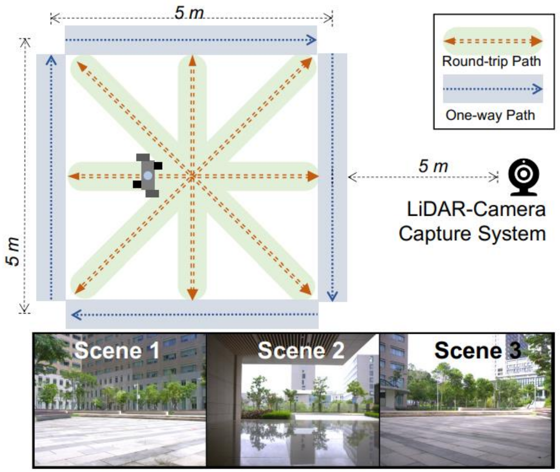



- Extended from 2D image-based Palette [Saharia+, CVPR'22]
  - Denoiser: 3D UNet with Relative Positional Embedding
- Initialization: $\mathbf{z}_t \leftarrow \mathbf{m} \odot \mathbf{y} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{z}_\mathrm{t}$
- Loss function: $\mathcal{L}_{T\to\infty} = \mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{1}),t\sim\mathcal{U}(0,1)}[\|\hat{\epsilon}(\text{concat}(\mathbf{y}, \mathbf{z}_t); \lambda_t) - \epsilon\|_2^2]$

# Part II / Experiments / Implementation Details

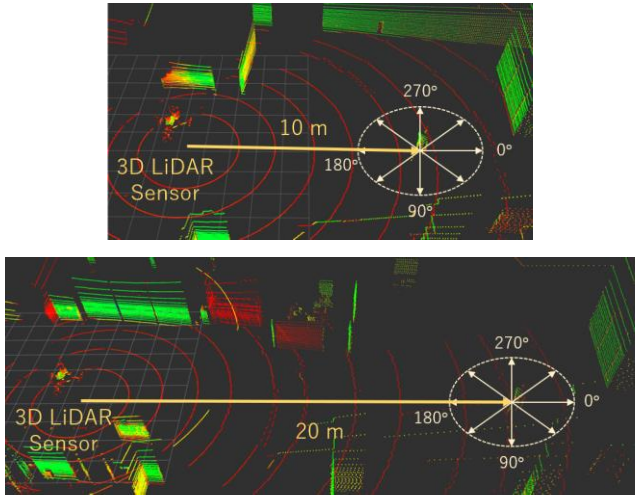
- Dataset comparison

SUSTeck1K [Shen+, CVPR'23]




For training and generalizability evaluation

KUGait30 [Ahn+, IEEE Access'23]



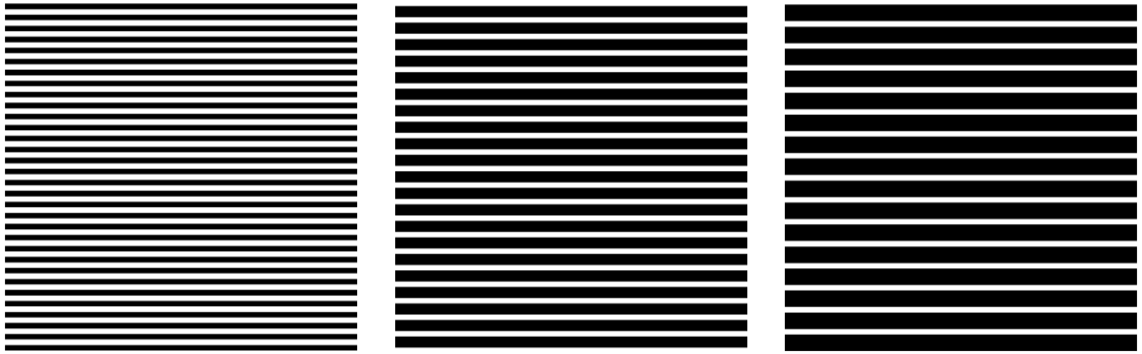

For practicality evaluation

| Datasets | Sensors | Beams | V/H Res. | Subjects | Angles | Distances |
|---|---|---|---|---|---|---|
| SUSTeck1K | VLS-128 | 128 | 0.11°/0.1° | 1,050 | 8 | 7.5 m |
| KUGait30 | VLP-32C | 32 | 1.33°/0.1° | 30 | 8 | 10, 20 m |

# Part II / Experiments / Implementation Details

- Used **noise masks** 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

→ Artificially degrade the complete gait data from SUSTeck1K by applying the combination of two different mask types

# Part II / Experiments / Implementation Details

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

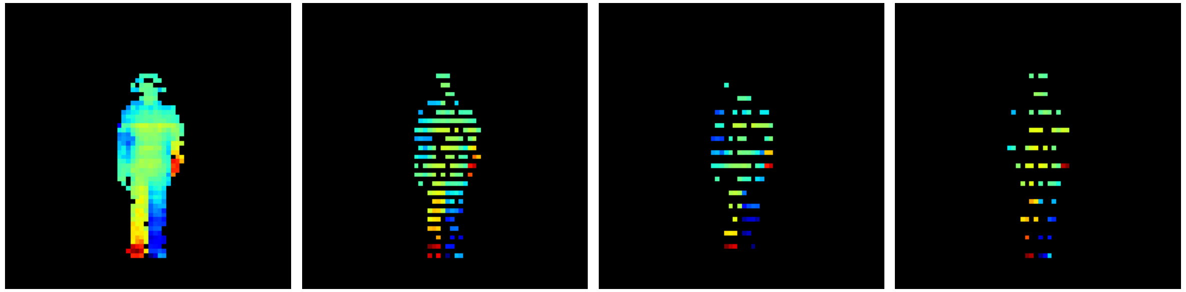
- Learning settings:
  - Learning rate: 0.0003
  - input sequence length: 10 frames
  - Timesteps: 32

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

- Experiments:
  - **Generative quality**:
    - Quantitative evaluation -> Qualitative evaluation
  - **Gait recognition task**:
    - Genealizability evaluation (on the SUSTeck1K) -> Practicality evaluation (on the KUGait30)

# Part II / Experiments / Generative Evaluation

- Quantitative results:
  - Compared methods:
    - Three interpolations
    - Palette
  - Metrics: PSNR, SSIM, Consistency



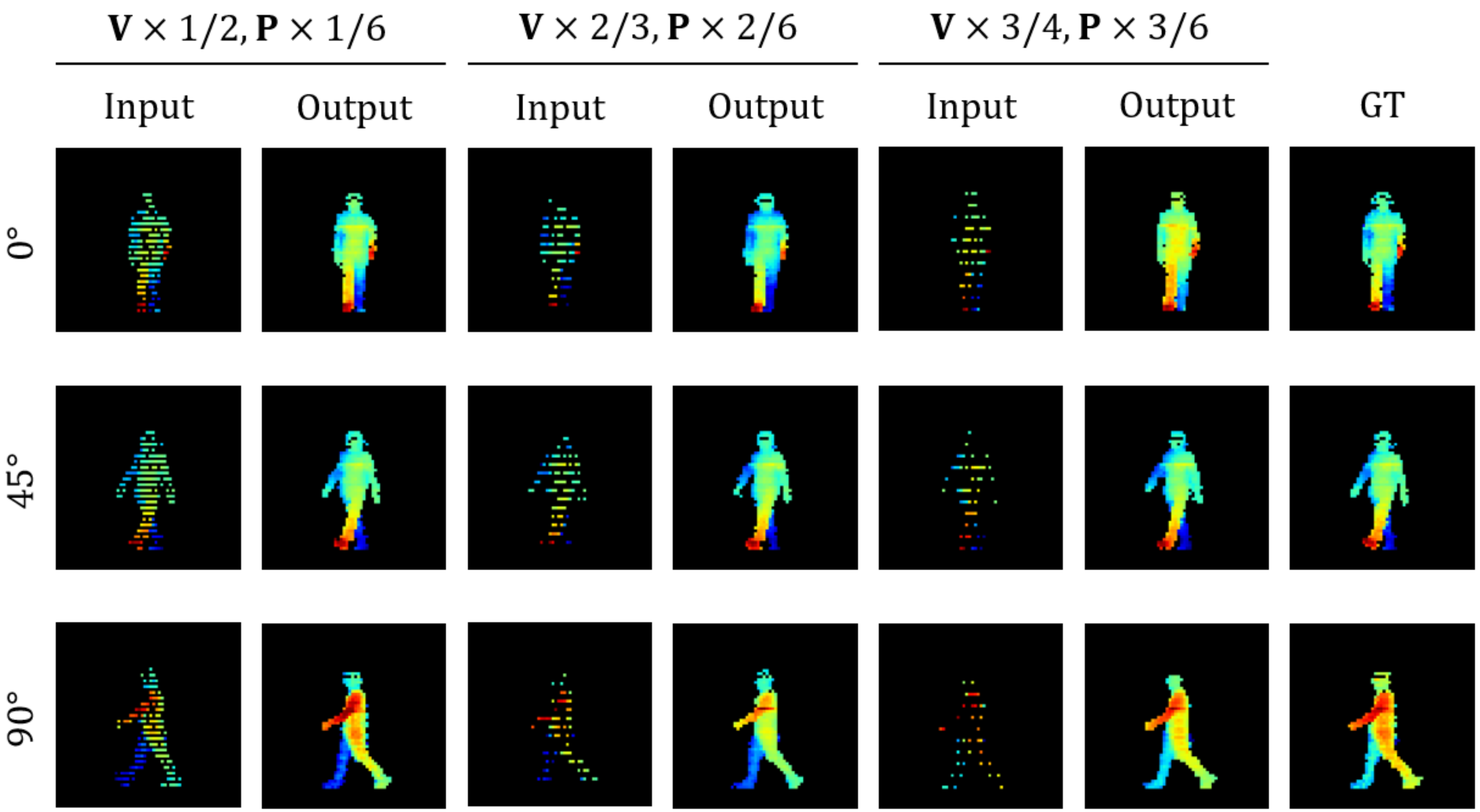
GT | **V** $\times 1/2$ + **P** $\times 1/6$ | **V** $\times 2/3$ + **P** $\times 2/6$ | **V** $\times 3/4$ + **P** $\times 3/6$

| | 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** |

→ Our model is superior to **all linear interpolations** and **vanilla Palette** across three metrics

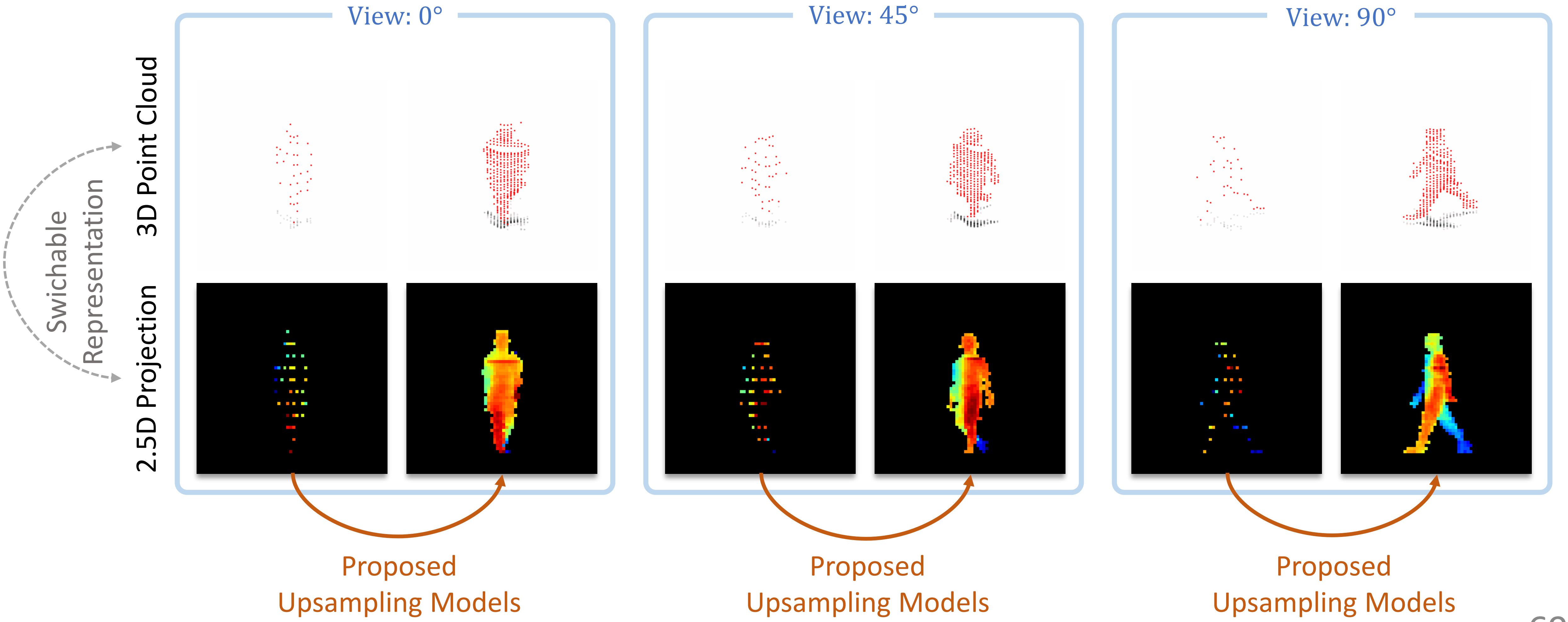
# Part II / Experiments / Generative Evaluation

- Upsampled results using the proposed model on SUSTeck1K



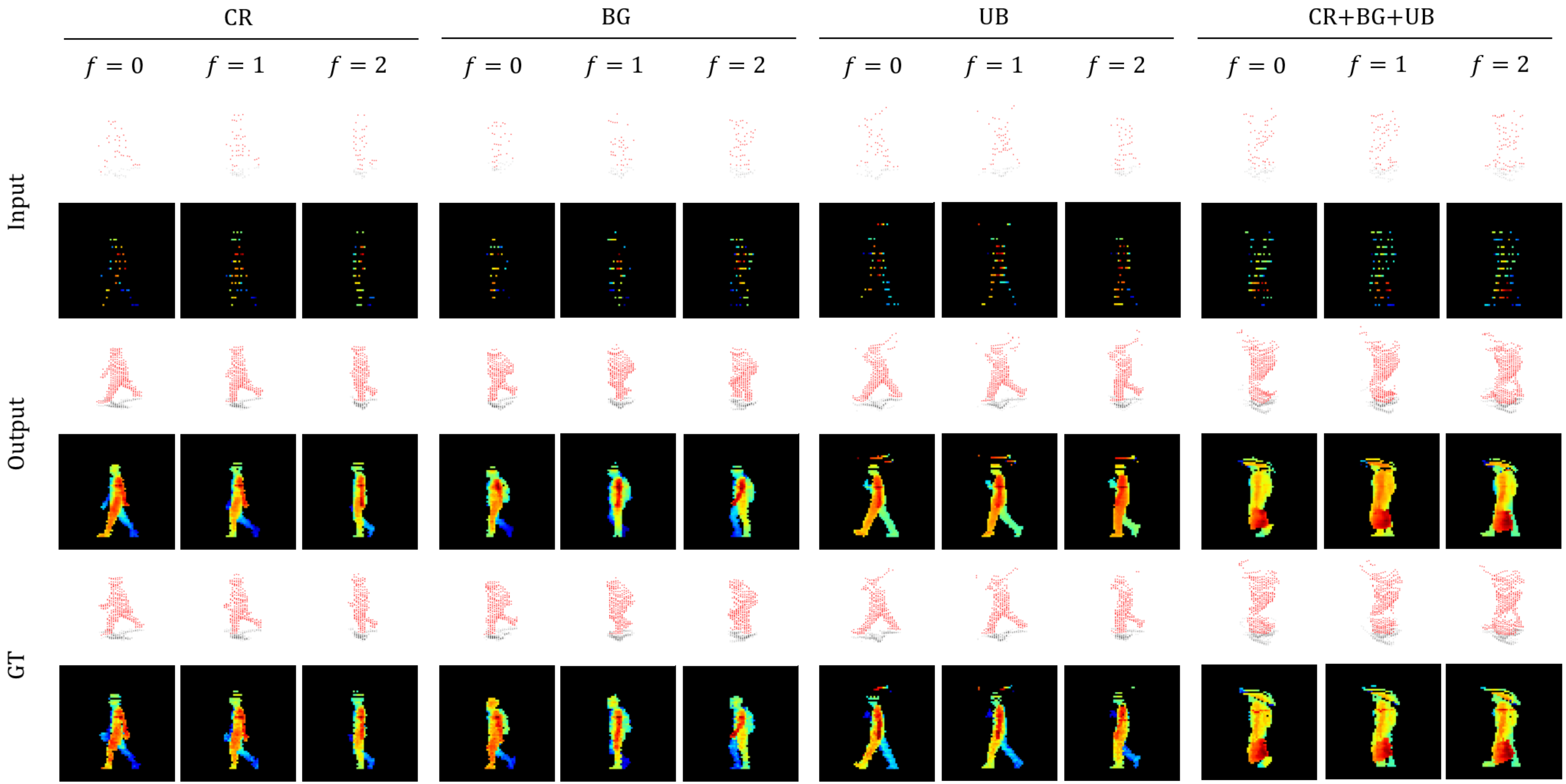
# Part II / Experiments / Generative Evaluation

- Upsampled results using our model **across three angles** on SUSTeck1K



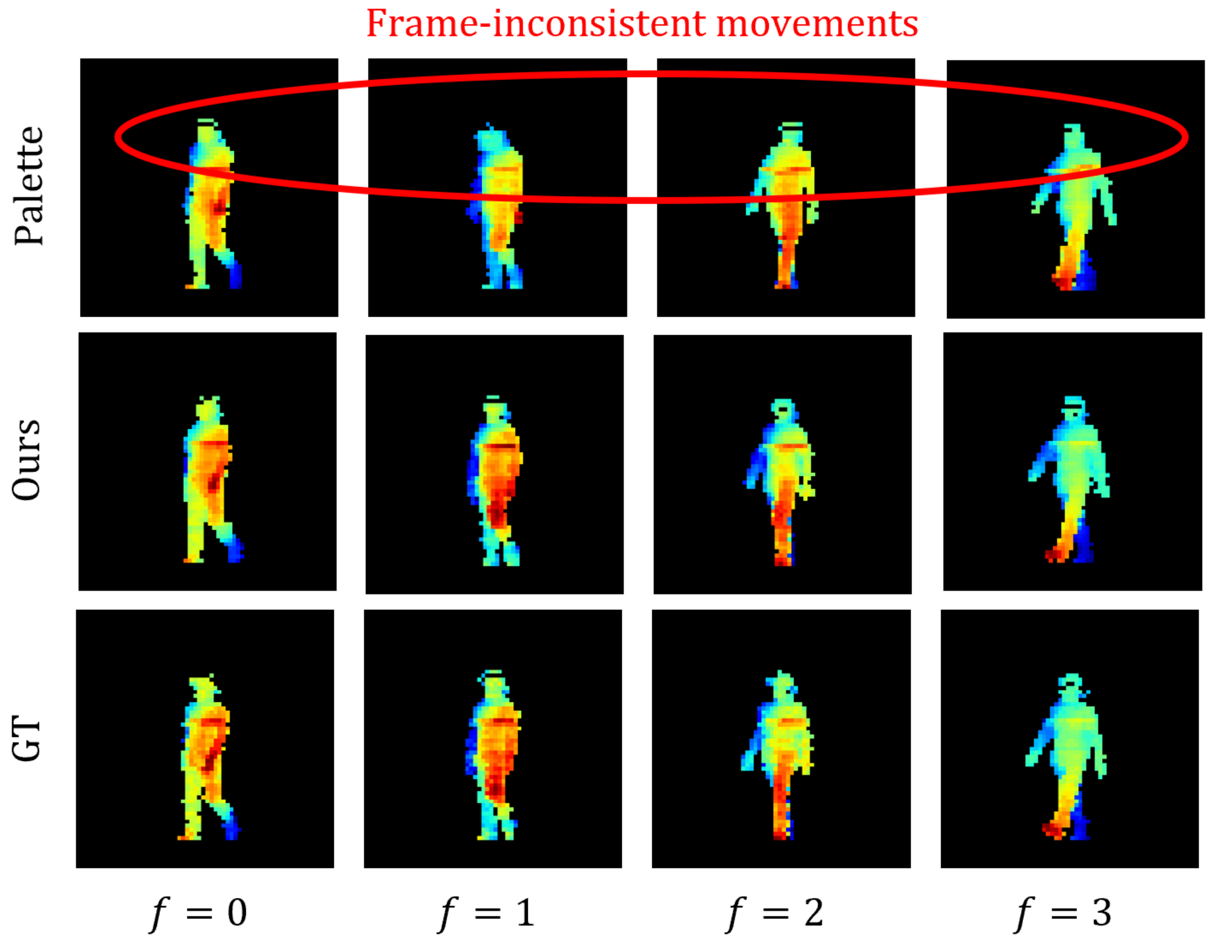
# Part II / Experiments / Generative Evaluation

- Upsampled results **with various attributes** using our model on SUSTeck1K



# Part II / Experiments / Generative Evaluation

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



# Part II / Experiments / Gait Recognition Task

- Quantitative results:
  - After **restoring missing parts in input data** with methods, **gait features are extracted** from the data by using the pre-trained LidarGait
  - Matche subject ID between **Gallery** and **Probe** by using **k Nearest Neighbor (kNN)**

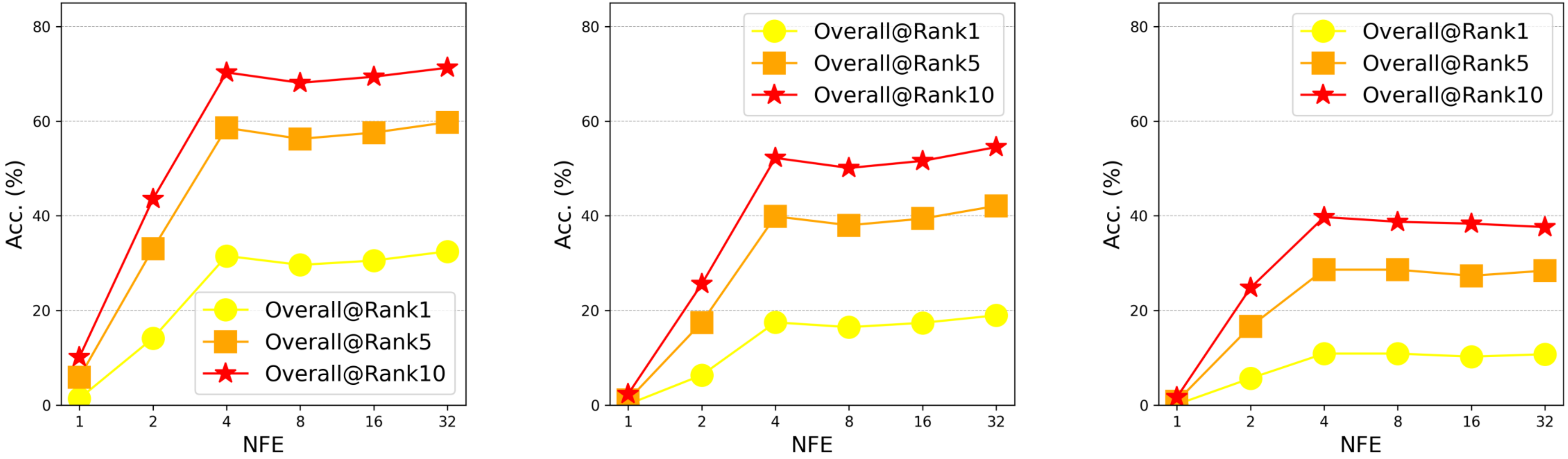
| Upsampling | | | Means (Probe set) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **V**×1/2, **P**×1/6 | | | **V**×2/3, **P**×2/6 | | | **V**×3/4, **P**×3/6 | | |
| Approach | Method | Input Modality | Rank1 ↑ | Rank5 ↑ | Rank10 ↑ | Rank1 ↑ | Rank5 ↑ | Rank10 ↑ | Rank1 ↑ | 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** |

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

# Part II / Experiments / Gait Recognition Task

- Comparison of the variations of timesteps for our model

NFE: Number of Function Evaluation

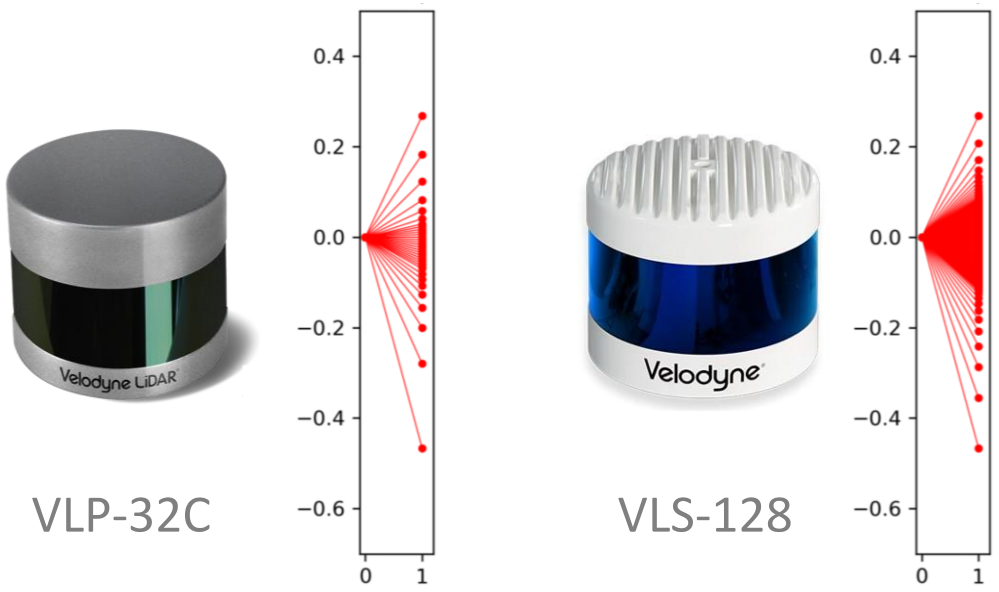



$\mathbf{V} \times 1/2 + \mathbf{P} \times 1/6$ $\quad$ $\mathbf{V} \times 2/3 + \mathbf{P} \times 2/6$ $\quad$ $\mathbf{V} \times 3/4 + \mathbf{P} \times 3/6$

→ The performance remains stable when the timestep is reduced to 4

# Part II / Experiments / Practicality

- Quantitative results
  - Training set: **SUSTeck1K** with noise masks (with **VLS-128**)
  - Testing set: **KUGait30** (with **VLP-32C**)
  - Significantly improve identification performance **even in real-world scenarios**

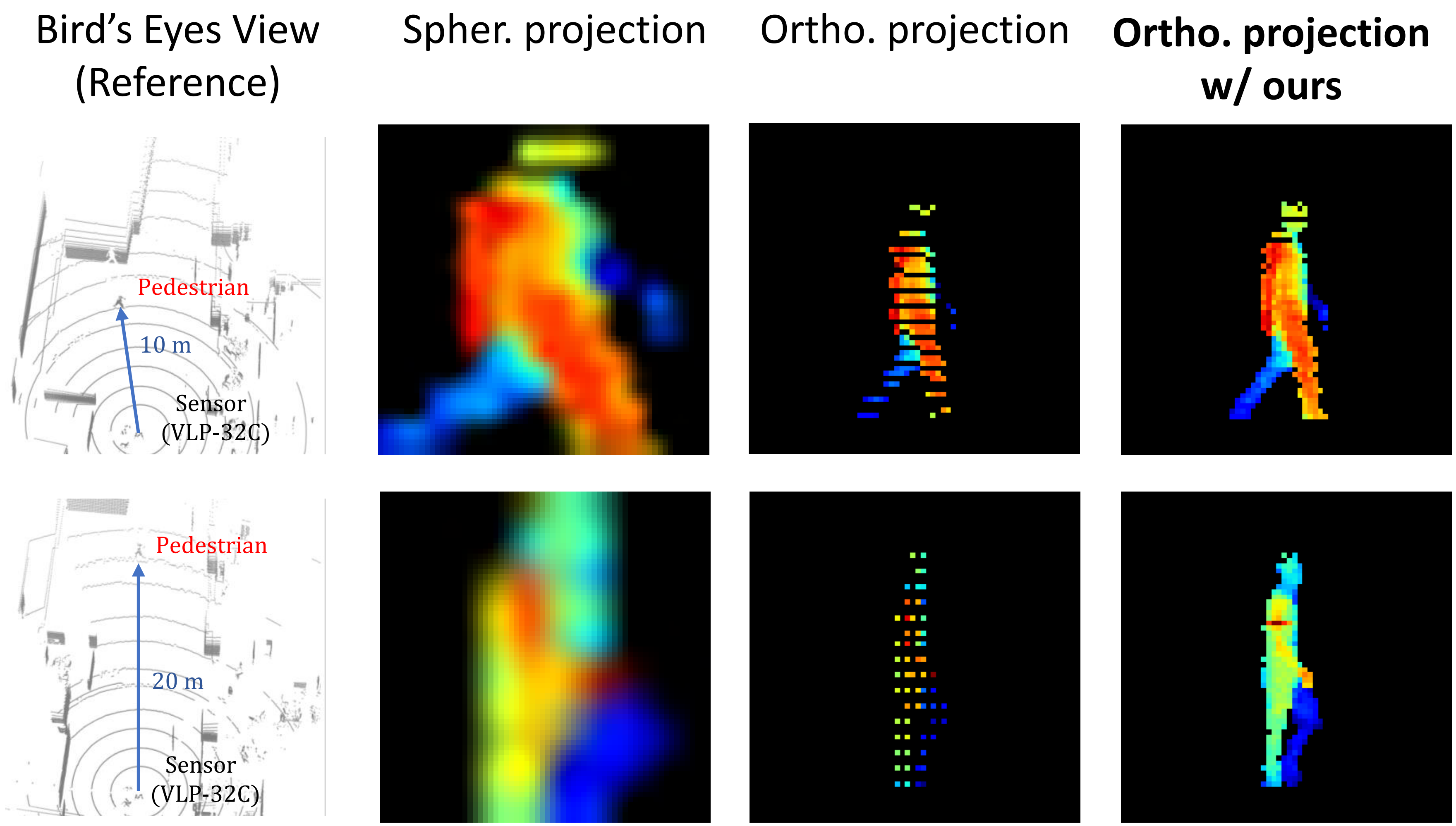



Angular resolution comparison

| 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** |

# Part II / Experiments / Practicality

- Qualitative results

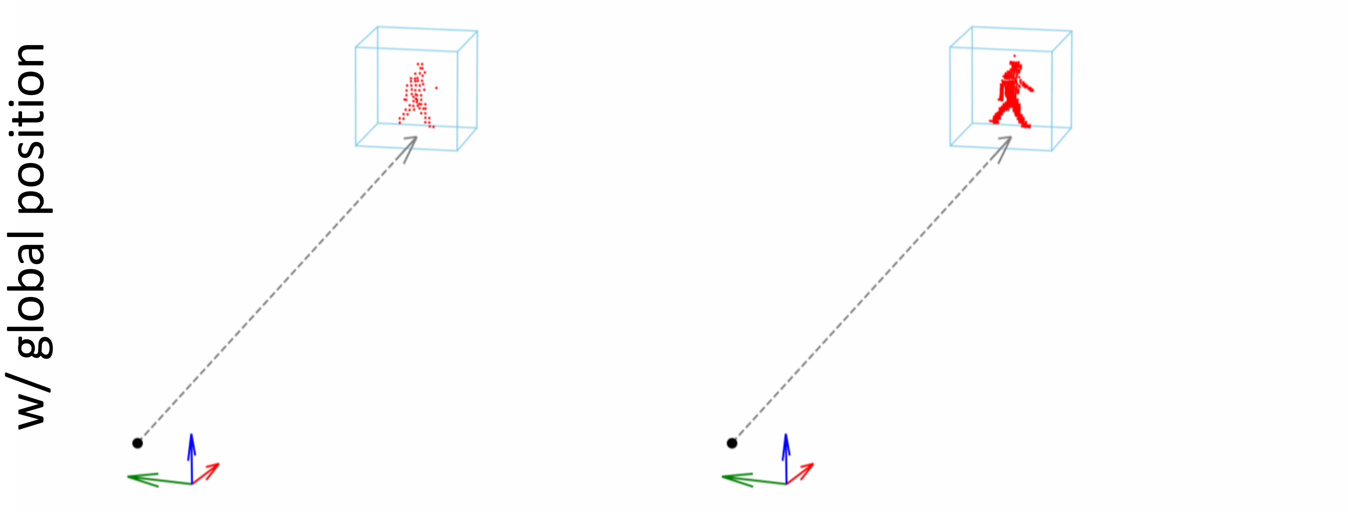


# Part II / Experiments / Practicality

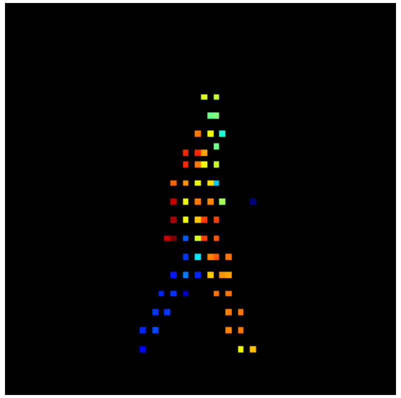
- Qualitative results

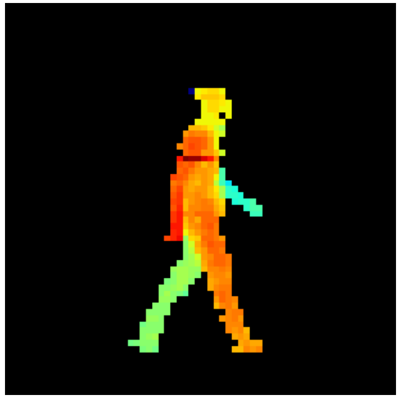
Original data (20 m) | Restored data

3D Point Cloud w/ global position



2.5D Projection





# Part II / Summary

- Introduced an upsampling model for LiDAR-based gait sequence data to address missing parts of walking shapes as an inpainting problem
- Demonstrated significant improvements in terms of both generative quality and identification performance
- Confirmed the effectiveness even for varying sensor type or measurement distance in real-world senarios

# Conclusion

- **Part I (Development of gait recognition models using 3D LiDAR):**
  - Reduces errors caused by linear interpolation by using orthographic projection
  - Enhances discriminative capability by leveraging the characteristics of LiDAR sensors

- **Part II (Development of gait upsampling models for 3D LiDAR):**
  - Improves the generalizability of identification models for long-distance
  - Addresses missing part of gait shapes as an inpainting problem

- Outlook
  - Task-agnostic approaches for more diverse real-world scenarios (including obstacle occlusion)
  - Consider employing Flow Matching (FM) to reduce inference speed

# Thank you for your attention!