3D LiDAR-based gait analysis for person identification in long-range measurement environments

A DISSERTATION
SUBMITTED TO THE GRADUATE SCHOOL OF INFORMATION SCIENCE
AND ELECTRICAL ENGINEERING
OF KYUSHU UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Jeongho Ahn
January 2025

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Ryo Kurazume)　Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Ryoma Bise)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Qi An)

Approved for the Kyushu University Committee on Graduate Studies

Abstract

Gait recognition is a biometric technology that identifies individuals based on their walking patterns. In particular, compared to other modalities such as face, fingerprint, or retina, it does not require user cooperation and allows for measurement from a long distance, making it highly promising for various person identification applications. With advances in deep neural networks, research on gait recognition using RGB cameras has made remarkable progress. However, challenges persist, including sensitivity to changes in camera angles and variations in lighting conditions, which can degrade identification performance. Recently, a 3D LiDAR sensor, which scans surrounding environments to acquire accurate three-dimensional data without being affected by lighting conditions, has been widely utilized in fields such as mobile robotics and autonomous driving. However, its application in gait recognition tasks remains limited. The primary challenges include its lower spatial resolution compared to conventional cameras and the inherent characteristics of 3D LiDAR, where laser beams diffuse spherically with increasing measurement distance, resulting in sparse pedestrian visual data.

To address these challenges, this dissertation proposes deep learning-based solutions to enhance identification performance for sparse gait data acquired from 3D LiDAR sensors in long-range environments. The dissertation is organized into two parts. In Part I, I present an identification method based on LiDAR projection that utilizes gait features from multiple viewpoints. From 3D LiDAR data, a pedestrian's trajectory is obtained in the global coordinate system, allowing for the estimation of walking direction. This makes it possible to generate invariant gait shapes, such as side and back views, which predominantly represent gait dynamics. Additionally, I propose several techniques for this identification method, such as multi-scale resolution encoding and attention-based feature fusion, to enable the effective learning of gait features under varying measurement distance scenarios, thereby maximizing the potential of 3D LiDAR. In Part II, I present an upsampling model for sparse gait sequence data. Sparse LiDAR gait data is addressed using orthogonal projection, framing it as an inpainting, which is a type of linear inverse problem. Additionally, a conditional diffusion model is employed to solve this simplified problem, with its effectiveness demonstrated through comprehensive experiments on both large-scale and real-world datasets. Taken together, these models provide distinct advantages across different aspects of LiDAR-based gait recognition.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor, Ryo Kurazume, for his invaluable guidance and unwavering support throughout my six-year journey during both my Master's and Ph.D. studies. He consistently encouraged and supported me in pursuing my research interests, even during challenging and uncertain times. He is also the most passionate and exceptional person who I have encountered in my research career, and I am truly fortunate to have had him as my mentor.

I would like to thank Yumi Iwashita for her invaluable support and advice as a sub-supervisor. She consistently provided dedicated guidance throughout my academic career, including my internship at NASA/JPL. From her, I have learned how to become an independent and passionate researcher.

I am deeply grateful to Kazuto Nakashima for his unwavering support and guidance as the mentor closest to my field of study. I would not have become who I am today without his help, as he guided me into the field of computer vision and supported me in becoming a creative and analytical researcher. From him, I have learned the importance of perseverance and have drawn continuous inspiration throughout my research journey.

I would also like to express my heartfelt thanks to Akihiko Kawamura, Shoko Miyauchi, and Kohei Matsumoto for their invaluable support and guidance throughout my studies. Their insights and encouragement have significantly inspired and shaped the direction of my research.

I feel fortunate to have had Koki Yoshino and Tomoya Itsuka as my Ph.D. colleagues, who greatly enriched my life at Fukuoka. They were always willing to share their wisdom and life experiences, for which I am truly grateful.

Finally, I would like to extend my deepest gratitude to my family, friends, and lab mates for their constant help and dedication, which made this dissertation and my life in Japan a truly memorable experience. The words written here are not enough to fully express my gratitude.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Prologue

In modern society, where the population continues to grow, automating the identification of large numbers of individuals has become a highly significant challenge. With decades of advancements in artificial intelligence (AI) technology, biometric person recognition, which leverages individuals' physiological characteristics, has made remarkable progress. For example, with the development of computer vision, various biometric modalities, such as iris, face and fingerprint, have made substantial advancements [11, 15, 56]. Among these, gait, which represents individuals' walking patterns, has attracted significant attention as a powerful tool in person identification tasks. Unlike other modalities, gait encompasses unique and dynamic physical as well as behavioral characteristics. Moreover, it offers the advantage of identifying individuals from a distance without requiring their cooperation or contact. In addition, it is notable for being non-intrusive and challenging to disguise. These distinct advantages make gait recognition particularly well-suited for applications in security systems and criminal investigations [7, 32, 46]. In the medical field, gait patterns are also used to diagnose or predict neurodegenerative diseases such as Alzheimer's, underscoring the importance of gait as a crucial factor in various applications [50, 85].

The workflow of the gait recognition system is illustrated in Fig. 1.1. Most gait recognition studies have predominantly used conventional RGB cameras to capture pedestrians [57]. These cameras have been adopted as visual sensors for gait recognition due to their affordability, ease of use, and high spatial resolution. Although significant progress has been achieved with these cameras, several challenging issues remain. Notably, varying lighting conditions caused by weather, climate, or background contrast often lead to failures in accurately segmenting pedestrians. Furthermore, pedestrian shapes can vary significantly depending on the camera's height, and linear interpolation may neglect the pedestrian's actual size, potentially reducing identification performance. To address these issues, some studies have explored the use of RGB-D cameras in the field of gait recognition,

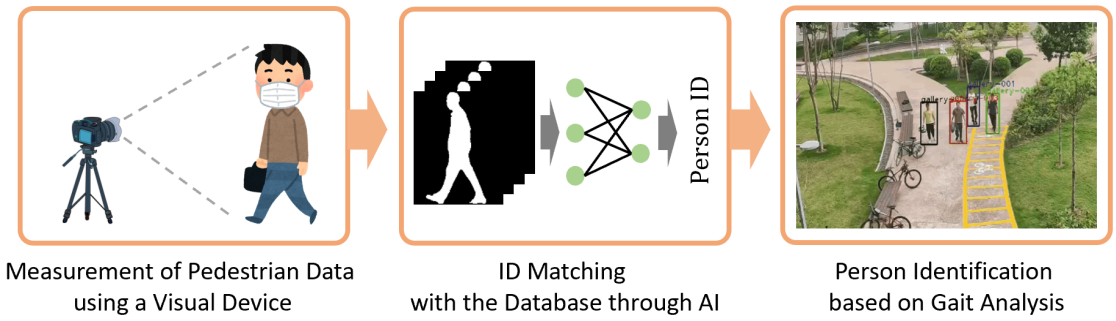


Figure 1.1: An illustration of a typical gait recognition system.

which can capture depth information [10, 80]. However, RGB-D cameras have been reported to have hardware limitations, such as a short measurement range and susceptibility to environmental noise, making them more suitable for indoor applications.

For decades, light detection and ranging (LiDAR) technology has played a pivotal role in the fields of computer vision and robotics, enabling the scanning of surrounding environments in 3D geometric dimensions. LiDAR operates by emitting laser pulses toward an object and measuring the distance based on the time it takes for the reflected light to return, as shown in Fig. 1.3 Compared to typical RGB cameras, LiDAR sensors can collect precise geometric information as 3D point clouds. Additionally, unlike RGB-D cameras, LiDAR sensors are more robust under various lighting conditions and can measure longer distances due to their active sensing based on short-wavelength pulsed lasers, making them well-suited for outdoor applications, such as mobile robots and self-driving vehicles [37, 71].

LiDAR sensors have been used primarily for scene understanding tasks in mobile robotics, such as object detection, tracking, or segmentation [44, 70, 81]; however, they have rarely been applied as tools in biometric systems, such as gait recognition. One potential reason is the relatively low spatial resolution of LiDAR sensors compared to general cameras, which makes it challenging to capture the detailed shapes of pedestrians. Some studies have shown that acceptable performance is achievable when identifying individuals at close range [6, 79]. However, due to the inherent characteristics of most LiDAR sensors, the laser pulses emit in a spherical pattern, causing the 3D point cloud data to become sparser as the measurement distance increases, as shown in Fig. 1.3. Furthermore, the resolution of LiDAR sensors also affects the sparsity of gait data, which is a key factor that could reduce the identification performance.

Despite these issues, LiDAR sensors hold significant potential in the field of gait recognition due to their ability to accurately capture pedestrian data in long-range 3D environments, even under challenging conditions such as low light or adverse weather. This capability is expected to overcome the limitations associated with typical RGB camera-based gait recognition. Furthermore, as the

(a) RGB camera (b) Depth camera (c) 3D LiDAR sensor

Figure 1.2: Visualization comparison among three visual sensors.

performance of LiDAR sensors improves and their costs decrease over time, their adoption as visual sensors is expected to increase rapidly. In particular, in mobile robotics and autonomous driving, where LiDAR technology is essential, outdoor person identification is anticipated to enable a variety of functions, including individual identification as well as applications such as surveillance robots and healthcare monitoring. Furthermore, LiDAR is more suitable for protecting individual privacy compared to general cameras, which capture explicit visual data of an individual's appearance.

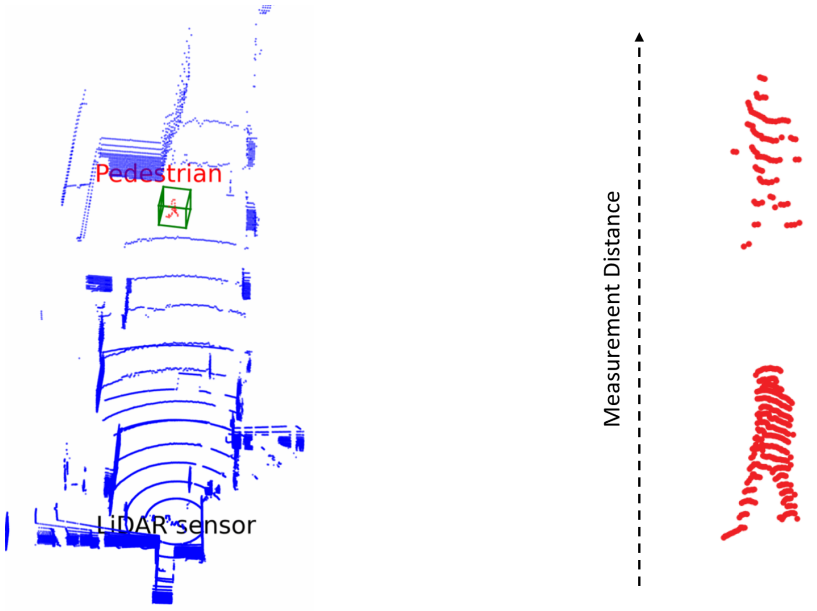
1.2 Goal

Based on the above, the primary challenge in utilizing LiDAR technology for practical gait recognition applications in outdoor environments lies in the sparse point cloud data resulting from long measurement distances. Motivated by this issue, my aim is to develop a gait recognition system using 3D LiDAR sensors in long-range measurement environments. Specifically, I propose models to enhance person identification performance using deep learning, a type of machine learning technique. Deep learning, as a data-driven approach, is well-known for its effectiveness in analyzing high-dimensional data. Owing to its power and flexibility, deep learning has found significant applications across various fields. In particular, it has greatly advanced the field of camera-based gait recognition [57] due to its capacity to uncover the complex and dynamic walking patterns. In this dissertation, the study is organized into two thematic parts: (1) the development of gait recognition models and (2) the development of gait upsampling models.

1.2.1 Study 1: The Development of Gait Recognition Models

The first study involves developing gait recognition models based on LiDAR projection representations. Data captured from 3D LiDAR sensors can generally be represented in two forms: (1) 3D point clouds and (2) range images, as illustrated in Fig. 1.4.

In LiDAR point clouds, each point represents xyz coordinates along with additional information such as intensity. In contrast, range images conventionally quantize each point through spherical

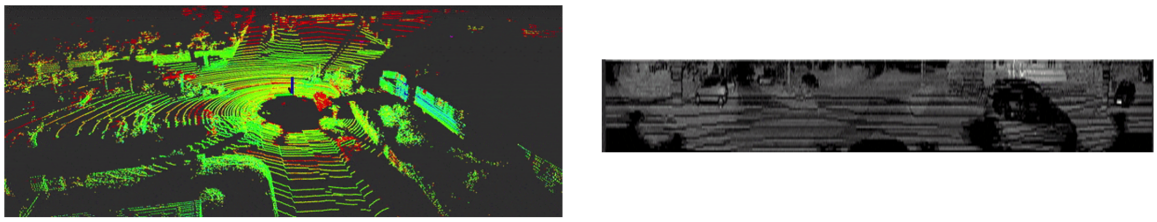


(a) Pedestrian detection obtained from a LiDAR sensor (b) Sparsity variations in LiDAR gait data

Figure 1.3: Comparison of pedestrian point cloud sparsity according to the measurement distance.

projection, which is determined by the hardware specifications of standard LiDAR sensors. Although the former ideally represents raw data obtained directly from LiDAR sensors and captures fine-grained patterns, it is often challenging to process due to its unordered nature and the time-consuming computations required. On the other hand, while the latter may introduce artificial errors during the quantization of 3D data into 2D grids, it is empirically favored in practical applications, such as scene understanding for mobile robotics, due to its fast processing speed and effectiveness in deep learning-related tasks [45, 76–78, 81]. Building on the effectiveness of projection-based approaches in LiDAR-related tasks, I develope a gait recognition model using a LiDAR projection strategy.

In a typical gait recognition system, pedestrians in images or videos captured by cameras are segmented and aligned to a uniform height using linear interpolation before being input into identification models. However, with LiDAR images generated through spherical projection, the inherently low spatial resolution of LiDAR sensors causes gait shapes to shrink significantly during long-range measurements, increasing quantization errors in the linear interpolation process. To address this challenge, I utilize orthogonal projection to represent gait features for the proposed recognition models. Unlike spherical projection, orthogonal projection quantizes the point cloud data in xyz space, allowing it to accurately reflect the actual size of pedestrians in gait images. Additionally, it preserves the overall shape of pedestrians, even as the data becomes sparser with increasing captured distance.



(a) 3D point cloud data (b) 2D spherical projection range image

Figure 1.4: Comparison of 3D LiDAR representations

To fully utilize the unique characteristics of LiDAR sensors, which are absent in conventional cameras, gait features are extracted from multiple viewpoints using orthogonal projection. In LiDAR data, a pedestrian's gait trajectory can generally be obtained through the global coordinate system when the LiDAR sensor is stationary. Leveraging this characteristic, I estimate the walking direction from the trajectory and project the gait shapes from fixed angles relative to it—specifically, the left-side and back views—to facilitate identification.

Despite the benefits of orthogonal projection which preserves the size of a target, the challenge still lies in handling sparse pedestrian shapes captured from long distances. To address this challenge, I propose a multi-scale spatial encoding strategy, which alleviates the impact of sparsity changes. Furthermore, in an extended study, an attention block is incorporated into this encoding module to determine the importance of each scale level based on the sparsity of gait shapes.

1.2.2 Study 2: The Development of Gait Upsampling Models

Recently, research on gait-based person identification using 3D LiDAR has started to gain momentum, particularly following the release of the large-scale LiDAR gait dataset SUSTeck1K [58]. However, relying solely on identification models to recognize sparse gait data poses challenges in fully understanding the underlying structure of gait appearance. Given that the representation of LiDAR data varies across identification models, it is particularly important to restore sparse raw gait data into complete data that resembles data captured at close range. To address this challenge, I develop a upsampling model designed for the sparse-to-dense restoration of LiDAR gait sequence data.

In sparse-to-dense approaches related to LiDAR, spherical projection is commonly used to restore low-resolution data into high-resolution images, in accordance with the hardware specifications of 3D LiDAR sensors [19, 47, 48]. However, this projection type targets the entire scene scanned by the LiDAR sensor, and the size of objects within the range image varies depending on the measurement distance. In contrast, the aforementioned orthogonal projection maintains the overall shapes of objects regardless of changes in the measurement distance. Particularly for long-distance measurements, the empty pixel regions within the pedestrian can be simplified using inpainting, a

type of linear inverse problem. Based on this assumption, I leverage orthogonal projection in the development of an upsampling model for LiDAR gait data. To implement this model, I employ diffusion probabilistic models (DPMs) [20, 27, 65], which sequentially corrupt training data by gradually adding noise and then learn to reverse this corruption in order to form a generative model of the data. DPMs have recently gained significant attention due to their superior diversity and high fidelity compared to other generative models, such as variational autoencoders (VAEs) [28] or generative adversarial networks (GANs) [14]. In particular, they have emerged as a key technology not only for image/video generation [53] but also for solving inverse problem tasks [43, 54, 63].

1.3 Dissertation Structure

This dissertation is comprised of two thematic parts.

Part 1 focuses on gait-based identification models parameterized by deep neural networks, aiming to enhance identification performance for sparse gait data captured over long distances.

- In Chapter 2, I present an identification model that incorporates invariant two-viewpoint gait features along with walking speed as inputs. Using the collected dataset, experiments are conducted to evaluate the robustness of the proposed model to variations in walking direction and the sparsity of gait data. This chapter is based on [1].
- In Chapter 3, I modify the identification model from Chapter 2 by introducing an attention-based module that adaptively learns gait features. Using a more comprehensive version of the collected dataset, experiments are conducted to evaluate the effectiveness of the proposed model. This chapter is built upon [2].

Part 2 is about developing an upsampling model for sparse LiDAR gait sequence data using diffusion probabilistic models.

- In Chapter 4, I present a conditional diffusion model for restoring LiDAR gait data using an orthogonal projection strategy. Additionally, I demonstrate the effectiveness of the proposed model through comprehensive experiments. This chapter is based on [3].

The discussion and future research avenues are concluded in Chapter 5.

Part I: The Development of Gait Recognition Models

Chapter 2

Identification Modeling for Range Variations

2.1 Introduction

RGB cameras dominate as visual sensors in the field of gait recognition [9, 12], and several previous studies have attempted to use depth or range information captured by RGB-D cameras (also known as depth cameras), such as the Microsoft Kinect [10, 80]. Compared to conventional cameras, RGB-D cameras offer several advantages, such as 3D information and simplified background subtraction. However, their measurement range and field of view are limited to approximately 10 m and 70° respectively, making them challenging to apply in outdoor environments. In recent years, 3D LiDAR sensors, capable of obtaining 3D range data of targets at distances exceeding 100 meters, have gained significant prominence. Especially, spinning 3D LiDAR sensors, a common type of 3D LiDAR, rotate 360 degrees to scan their surroundings and generate comprehensive 3D point clouds of the environment. Thanks to these characteristics, they have garnered significant attention in computer vision and have been widely applied in autonomous robots, acting as eyes for tasks such as object detection and navigation [41, 81]. Compared to RGB and depth cameras, however, these LiDAR sensors have rarely been used in biometrics. One possible reason is their low resolution, which makes it challenging to capture complete human shapes. However, considering the unique characteristics of 3D LiDAR, such as robustness to illumination conditions, long measurement distances, and a 360° scanning range in the azimuth, it holds significant potential for outdoor applications as a biometric identifier. Furthermore, it can be used as an alternative to conventional cameras in terms of protecting personal information.

I previously proposed an identification model using 3D LiDAR sensors and demonstrated that 3D LiDAR has significant potential for recognition tasks [79]. However, in this study, both the

measurement distance and the walking direction from the 3D LiDAR sensor were constant. This scenario is applicable when the LiDAR sensor is placed in a corridor or narrow street, with people walking in a single direction. I anticipate that 3D LiDAR will be increasingly adopted for more practical applications in person identification systems in the future, involving complex scenarios such as long-range or wide-area person identification. Additionally, gait recognition using 3D LiDAR sensors can serve as a functionality for mobile robots and autonomous driving systems. For example, security robots, which can operate 24/7 and are less conspicuous than human personnel, are increasingly being deployed in malls, offices, and public spaces. A nighttime surveillance system can be achieved by applying biometrics to these security robots without the installation of other sensors. Moreover, autonomous vehicles can identify and track specific users while in operation. Given these applications, it is essential to design a robust gait recognition model that accounts for intra-subject variations, particularly changes in viewing angles and measurement distances.

Building on the aforementioned study [79], this chapter introduces a gait-based person identification model using a 3D LiDAR sensor, designed to be robust against variations in measurement range and walking direction. To reduce the influence of these variations, I focus on two fixed invariant viewpoints and the walking pace, which may be invariant gait features for the conditions above, as shown in Fig. 2.1. In contrast to previous studies [79] that employed spherical projection from the sensor's perspective, this model utilizes an orthogonal projection strategy to map 3D pedestrian point clouds onto 2D planes, accurately reflecting real-world sizes and enhancing discriminative capability.

The contributions in this chapter can be summarized as follows:

- I first attempt to develop a gait recognition system using 3D LiDAR that is robust to variations in walking direction and the measurement distance from the sensor.
- To enhance identification performance, I leverage the unique characteristics of 3D LiDAR, such as its ability to capture 3D spatial and positional information, which are absent in conventional cameras.

2.2 Related work

2.2.1 Gait Representation in RGB Cameras

Gait recognition approaches for RGB videos can be categorized into two types: model-based and appearance-based methods. These categories depend on how the gait-related features are designed for walking.

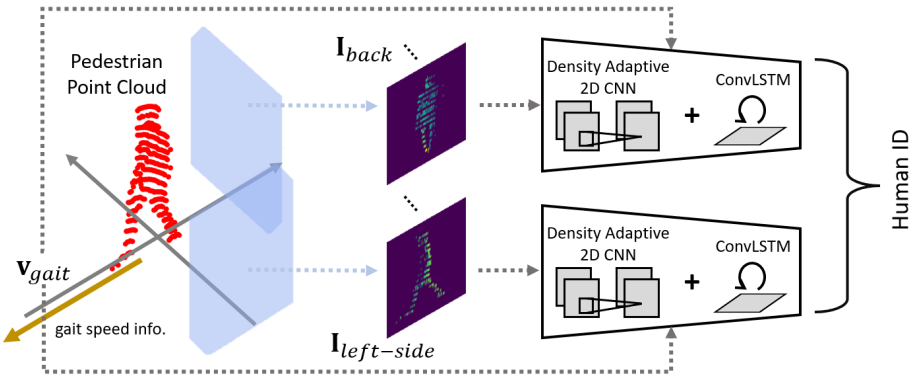


Figure 2.1: Overview of the proposed identification model, utilizing the unique characteristics of 3D LiDAR that are not found in regular RGB cameras. After estimating the walking direction of a subject point cloud, the two viewpoint-invariant gait image sequence and gait speed sequences are extracted and fed into the recognition network to identify an individual.

Model-based Methods

In model-based methods, first, the pose estimation algorithm is applied to model an articulated body with geometric properties, such as the lengths of skeletons, stride, cadence, and joint angles [4,67]. These approaches are generally immune to appearance changes because gait-related dynamics representing the joint information of the human body are learned under different clothing conditions. With the rapid development of human pose estimation methods, the recognition accuracy of model-based approaches has considerably improved [34,35,69]. In particular, studies [34,87] have achieved high recognition performance by adopting a 3D human representation [42] that includes not only pose but also shape parameters. However, these approaches remain challenging because of their heavy reliance on accurate key point estimation of image sequences, and sensitivity to occlusions, which could lead to the loss of identity-related shape information.

Appearance-based Methods

Compared with model-based approaches, shape-related gait features from original videos are directly used in appearance-based approaches. For example, a gait energy image (GEI) [17] is an appearance-based approach in which a silhouette sequence of the gait cycle is averaged to represent spatio–temporal information. Extended GEI-like modalities, such as frame difference frieze patterns [60], gait flow patterns [31], and affine moment invariants [24], have been proposed. Furthermore, the performance is improved by feeding GEIs into CNNs [61], and this approach has been used in other studies as a baseline. However, these methods compress time-series information into a single frame, which leads to a loss of opportunity for applying gait dynamics in temporal changes. In contrast, silhouette images, which describe body states in binary, have become popular in general gait

recognition tasks as input representation because of their effectiveness in recognition performance and low computational cost. For example, Chao *et al.* [9] achieved promising results by integrating gait silhouettes as a set, and [13] extracted spatio–temporal features from each body part. In contrast, Fan *et al.* [39] proposed a global–local convolution approach to address the neglect of local region details in gait frames, and a subsequent version [38] designed a cross-domain evaluation to synthesize both segmentation and recognition networks. Auto-encoders that disentangle appearance into style and pose features have been proposed to address gait-irrelevant variables (e.g., clothing and carrying) in RGB images [86]. The separation performance was further improved by augmenting the training data through adversarial generation [82]. I focused on the appearance-based approach assuming that inferring accurate key point locations remains challenging for gait recognition due to the sparseness and incompleteness in pedestrian point clouds captured from general LiDAR sensors, despite studies on LiDAR-based pose estimation [33].

2.2.2 Gait Recognition using Range Sensors

Compared with RGB cameras, few studies based on range-based sensors have been conducted for gait recognition. Kozlow *et al.* [30] classified the gait types of individuals using RGB-D cameras by using a 3D skeleton model with Bayesian networks, whereas Sadeghzadehyazdi *et al.* [51] applied flash LiDAR sensors with both 2D and 3D skeleton models. In studies utilizing LiDAR sensors, Benedek *et al.* proposed GEI-based methods [5, 6, 16] for re-identifying individuals within a short range and a limited time frame. However, this method cannot satisfactorily extract dynamic features from gait frames, which are critical for discriminating individuals. To address this problem, I proposed a method [79] for exploring temporal gait changes using LSTMs [21]. This study exploited depth representation in a spherical projection, which has been used in several LiDAR-related tasks for its processing efficiency [47, 48]. However, this method exhibits degraded recognition performance when the walking directions and distances measured from LiDAR sensors are not constant, which limits the flexibility of the model in real-world scenarios with complex confounding conditions.

2.3 Datasets

2.3.1 Data Collection

In this section, I describe the dataset used in subsequent experiments to verify the effectiveness of the proposed model. The dataset was captured using a 32-beam LiDAR scanner (Velodyne HDL-32E), which generates horizontal 360° range images. It consists of point cloud sequences from 31 individuals, including 28 men and 2 women. Data collection took place in the summer of 2020 at the West 2 building on the Kyushu University campus, with the LiDAR sensor operating at 10 frames per second (FPS).

During the data collection, participants were instructed to walk naturally along a circular path with a radius of 5 meters. The center of the circle was positioned 8.5 meters away from the sensor, as illustrated in Fig. 2.2a. As a result, the dataset captures variations in 360° walking directions and distances ranging from 3.5 to 13.5 meters. Compared to the dataset used in previous studies [79], this dataset introduces simultaneous changes in gait direction and data sparsity across frames, making it particularly challenging for recognition tasks.

2.3.2 Preprocessing

To evaluate the robustness of the proposed model to variations in walking direction and measurement distance, I constructed four datasets by segmenting the gait sequence data described above, as illustrated in Fig. 2.2b. Each dataset differs in terms of viewing angles and measurement distances. Let $\mathbf{P}_t$ be the subject point set for the time step t, extracted by background subtraction processing:

$$\mathbf{P}_t = \{\mathbf{p}_{t,1}, \mathbf{p}_{t,2}, ..., \mathbf{p}_{t,N}\}, \tag{2.1}$$

where N is the total number of points for each subject, and each point $\mathbf{p}_{t,n} \in \mathbb{R}^3$ indicates its orthogonal coordinates $(p_{t,n,x}, p_{t,n,y}, p_{t,n,z})$. Given a subject point cloud extracted, I define the center of gravity for a subject, $\mathbf{c}_t = (c_{t,x}, c_{t,y}, c_{t,z})$, as follows:

$$\mathbf{c}_t = \frac{1}{N}\sum_{n=1}^{N} \mathbf{p}_{t,n}, \tag{2.2}$$

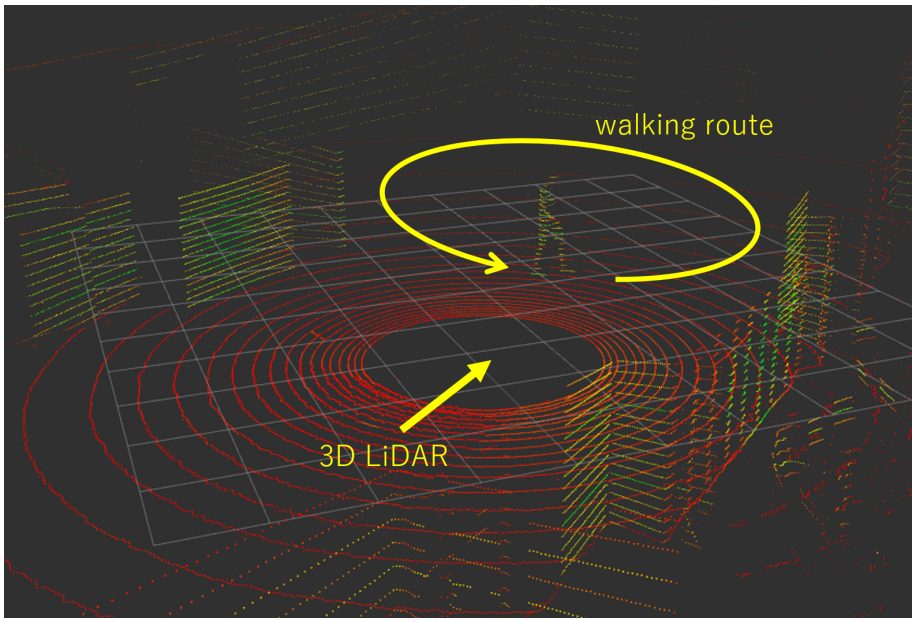
where $c_{t,z}$ was set to 0 to simplify the distance calculation. Given a subject central point $\mathbf{c}_t$, the distance from the sensor position $\mathbf{o}$ to a subject d_t can be calculated as follows:

$$d_t = \|\mathbf{c}_t - \mathbf{o}\|_2, \tag{2.3}$$

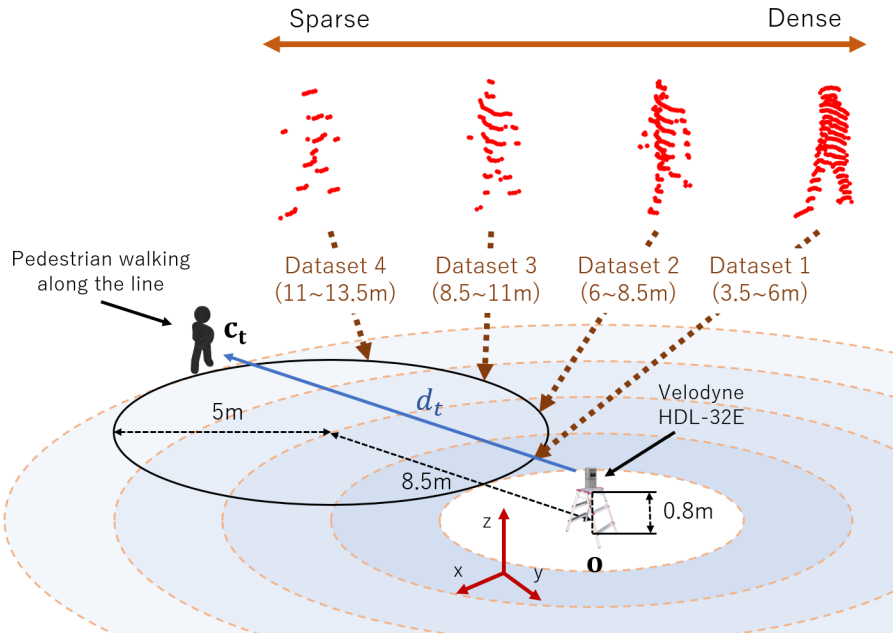
Finally, the extracted time-series gait data were divided into four subsets according to the calculated distance d_t: datasets 1, 2, 3, and 4 containing gait sequences with ranges of $3.5 - 6$ m, $6 - 8.5$ m, $8.5 - 11$ m, and $11 - 13.5$ m, respectively.

2.4 Method

In this section, I present a pipeline for the proposed gait recognition model, which consists of three components: gait direction transformation, input generation, and a recognition network. In particular, to enhance the robustness of changes in the walking direction and the distance measured from a sensor, I take advantage of 3D LiDAR, which is not included in normal RGB cameras.



(a) Overview of data collection



(b) Three-dimensional point cloud visualization

Figure 2.2: Data acquisition environment. The subjects walked along a circular line with a radius of 5 m, the center of which was 8.5 m away from the sensor. To evaluate the robustness of this approach, the gait data extracted through background subtraction were divided into four datasets according to the distance measured d_t.

2.4.1 Gait Direction Transformation

I first describe how to estimate the walking direction of a pedestrian and transform the subject point cloud into a new subject point cloud heading in a constant direction. In other words, the walking direction of a subject point set is transformed into the direction toward the $-y$-axis and generated into left-side view gait images from the yz coordinate plane. First, the gait directional angle for a time step t can be calculated from the subject's central points before and after a time step t:

$$\theta_t = \arctan \frac{c_{t+1,y} - c_{t-1,y}}{c_{t+1,x} - c_{t-1,x}} \tag{2.4}$$

Given a directional angle θ_t, a subject point cloud transformed $\hat{\mathbf{P}}_t = \{\hat{\mathbf{p}}_{t,1}, \hat{\mathbf{p}}_{t,2}, ..., \hat{\mathbf{p}}_{t,N}\}$ is obtained by rotating the original subject point cloud $\mathbf{P}_t$ around $\mathbf{c}_t$ as the z-axis.

$$\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_t - \pi/2) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t), \tag{2.5}$$

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

$$\mathbf{R}_z(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.6}$$

The case of generating a back-view gait image follows the same procedure, except that the rotation matrix $\mathbf{R}_z(-\theta_t - \pi/2)$ is replaced with $\mathbf{R}_z(-\theta_t - \pi)$.

2.4.2 Input Generation

In the next step, I describe how to generate the input data, which consist of a left-side view and back-view gait image sequences, and the gait speed sequences from the pedestrian point cloud transformed above for input into a recognition network. The gait image $\mathbf{i}_t \in \mathbb{R}^{V \times H \times 1}$ is generated from the pedestrian heading along the $-y$-axis:

$$\mathbf{i}_t = f_{\text{img}}\left(\hat{\mathbf{P}}_t\right), \tag{2.7}$$

where $f_{\text{img}}(\cdot) : \mathbb{R}^{N \times 3} \rightarrow \mathbb{R}^{V \times H \times 1}$ extracts a gait image representing the depth information of a pedestrian, as shown in Fig. 2.3. Here, the depth value and the size of the height channel are determined as $V(= l_z/l_{z\text{-res}}) \times H(= l_y/l_{y\text{-res}})$ pixels, where l_z, l_y, $l_{z\text{-res}}$, and $l_{y\text{-res}}$ are the height of the z-axis, the width of the y-axis, the vertical resolution of the image, and the horizontal resolution of the image, respectively. As the criteria by which each pixel value of the depth channel is determined, the value is set to 0 when none of the points correspond to the pixel. Otherwise, it is set to the largest x-coordinate value in the candidate set $\{\hat{\mathbf{p}}_{t,1}, \hat{\mathbf{p}}_{t,2}, ..., \hat{\mathbf{p}}_{t,\tilde{N}}\} \subset \hat{\mathbf{P}}_t$, which are the points corresponding

to the pixels in (v, h). In addition, for a clear distinction between a pedestrian and a background, the constant $l_x/2$ is added to the pixel value $i_{t,v,h}$ corresponding to

$$i_{t,v,h} = \begin{cases} \max\limits_{\tilde{n}\in\{1,\ldots,\tilde{N}\}} (\hat{\mathbf{P}}_{t,x}) + \frac{l_x}{2} & (\exists \hat{\mathbf{p}}_{t,\tilde{n}}) \\ 0 & (\text{otherwise}) \end{cases} \tag{2.8}$$

Here, the candidate point $\hat{\mathbf{p}}_{t,\tilde{n}}$ should satisfy the following formulation:

$$\hat{\mathbf{p}}_{t,\tilde{n}} = \{\hat{\mathbf{p}}_{t,\tilde{n}} \in \hat{\mathbf{P}}_t \mid (v = \tilde{v}) \wedge (h = \tilde{h})\}, \tag{2.9}$$

where $\tilde{v}$ and the $\tilde{h}$ are defined as follows:

$$\tilde{v} = \left\lfloor \frac{1}{l_{z\text{-res}}} \cdot \left(\hat{p}_{t,\tilde{n},z} + \frac{l_z}{2} \right) \right\rfloor, \tag{2.10}$$

$$\tilde{h} = \left\lfloor \frac{1}{l_{y\text{-res}}} \cdot \left(\hat{p}_{t,\tilde{n},y} + \frac{l_y}{2} \right) \right\rfloor \tag{2.11}$$

Thus, the gait image sequence $\mathbf{I} \in \mathbb{R}^{T\times V\times H\times 1}$ with T frames is obtained:

$$\mathbf{I} = (\mathbf{i}_1, \ldots, \mathbf{i}_T) \tag{2.12}$$

As another input for the recognition network, the gait speed can be obtained from the pedestrian central points before and after time step t and the rotation rate f_{rps} of the sensor:

$$v_t = \frac{f_{\text{rps}}}{2} \cdot \|\mathbf{c}_{t+1} - \mathbf{c}_{t-1}\|_2 \tag{2.13}$$

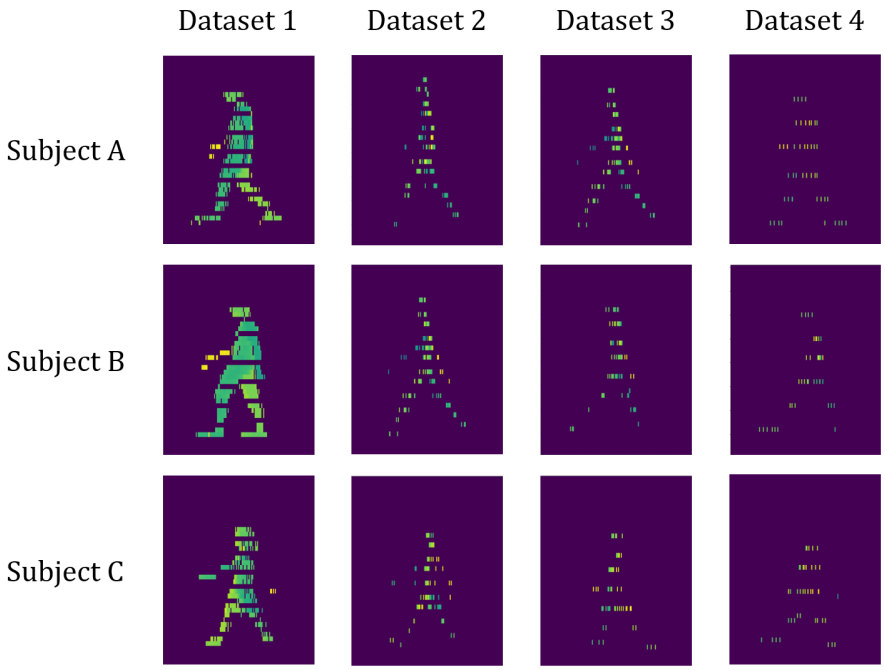
Based on the calculation above, we can obtain the gait speed sequence $\mathbf{v}_{\text{gait}} \in \mathbb{R}^T$ with T frames:

$$\mathbf{v}_{\text{gait}} = (v_1, \ldots, v_T) \tag{2.14}$$

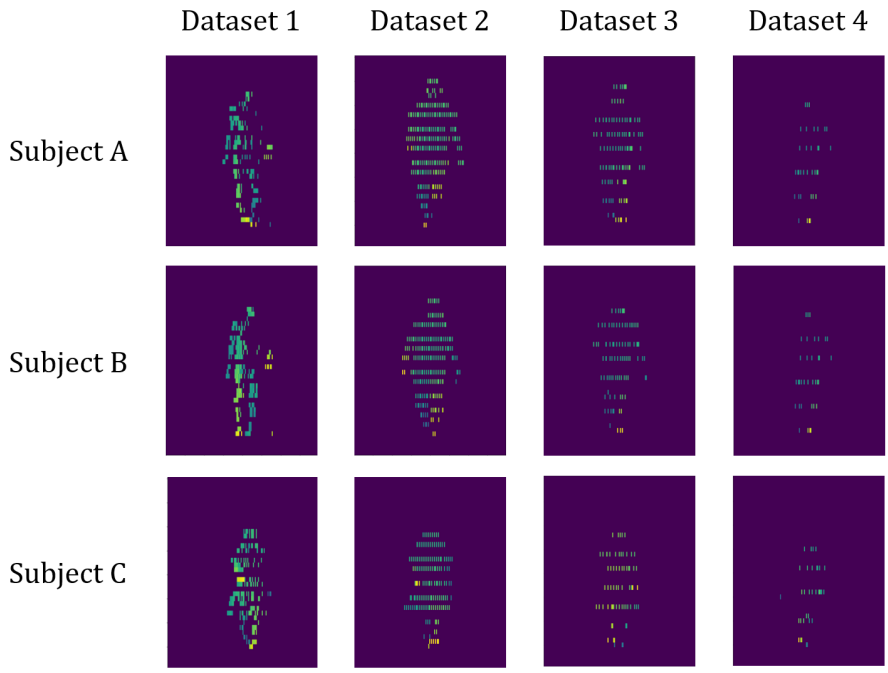
Finally, we can obtain the left-side view gait image sequence $\mathbf{I}_{\text{left-side}}$, the back-viewed gait image sequence $\mathbf{I}_{\text{back}}$, and the gait speed sequence $\mathbf{v}_{\text{gait}}$ for the proposed recognition network. In this study, l_z, l_y, $l_{z\text{-res}}$, $l_{y\text{-res}}$, T, and f_{rps} are set as 2.4 m, 1.6 m, 0.06 m, 0.01 m, 10, and 10 Hz, respectively.

2.4.3 Recognition Network

In this section, I describe the recognition network designed to learn the discriminative information from the inputs generated in the previous step. The overall architecture, which consists of four key components-density adaptive encoding (DAE), temporal feature aggregating (TFA), positional feature concatenating (PFC), and viewpoint-informed feature aggregating (VFA)-is illustrated in



(a) Left-side view



(b) Back view

Figure 2.3: Example of a generated gait image sequence, which is one of the inputs of the recognition network. The point cloud of the subject is sparse when the target is at a long distance.

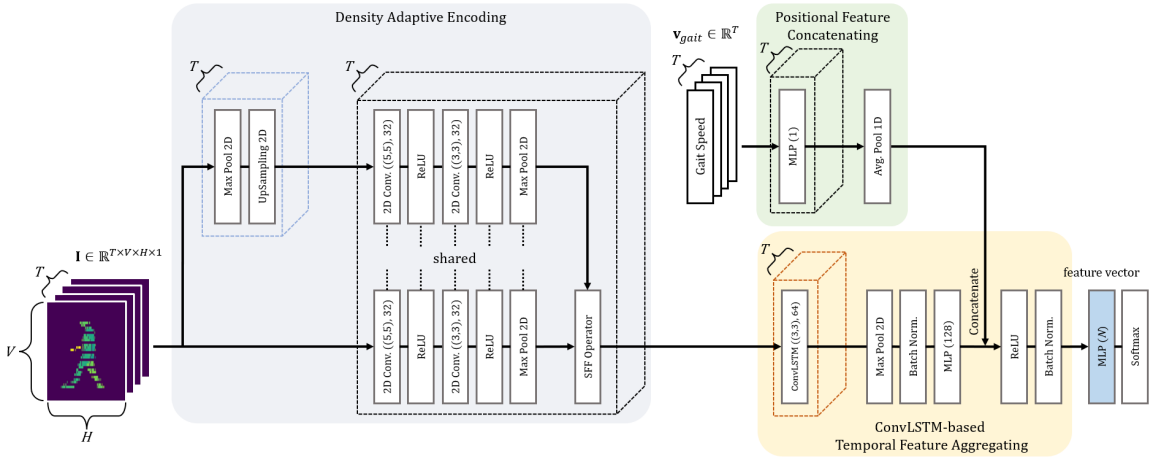


Figure 2.4: The overall architecture of the proposed recognition network. The gait image and speed sequence are fed into the network as inputs to learn the spatial-temporal and positional features and improve the discriminative capability of the gait recognition.

Fig. 2.4.

Density Adaptive Encoding

A high-resolution image, transformed from the point set of a subject, is expected to capture fine-grained patterns of the gait. However, such images lack the capability to recognize the entire human body in sparse data. Consequently, spatial features learned from dense data captured at short distances may not generalize well to long distances, as point sets typically exhibit non-uniform density at varying distances. To address this limitation, I designed a module called density adaptive encoding (DAE). This module leverages low-resolution representations, which are potentially more robust to sparse data and better suited for recognizing coarse-grained patterns, enabling the learning of multi-scale features by combining information extracted from different resolutions.

First, the double-reduced-resolution image sequence $\mathbf{I}_{\text{low-res}} \in \mathbb{R}^{T\times V/2\times H/2\times 1}$ is obtained by feeding a gait image sequence $\mathbf{I}$ into the max pooling layer:

$$\mathbf{I}_{\text{low-res}} = \text{Maxpool2D}(\mathbf{I}) \tag{2.15}$$

Then, $\mathbf{I}_{\text{low-res}}$ is upsampled to meet the height and width dimensions of $\mathbf{I}$:

$$\hat{\mathbf{I}}_{\text{low-res}} = \text{Upsampling2D}(\mathbf{I}_{\text{low-res}}) \tag{2.16}$$

Next, two gait image sequences with different resolutions, $\mathbf{I}$ and $\hat{\mathbf{I}}_{\text{low-res}}$, are fed into the 2-D convolution network module Conv2D$(\cdot)$, which shares the weights and kernel optimizations, and obtains the spatial features, $\mathbf{F}, \mathbf{F}_{\text{low-res}} \in \mathbb{R}^{T\times V/2\times H/2\times 16}$, respectively. Finally, $\mathbf{F}$ and $\mathbf{F}_{\text{low-res}}$ are combined

into one feature sequence $\hat{\mathbf{F}}$ using the spatial feature fusion (SFF) operator as follows:

$$\hat{\mathbf{F}} = \frac{1}{2} \cdot (\text{Conv2D}(\mathbf{I}) \ \oplus \ \text{Conv2D}(\hat{\mathbf{I}}_{\text{low-res}})), \tag{2.17}$$

where $\oplus$ represents an element-wise addition.

Temporal Feature Aggregating

The Temporal Feature Aggregating (TFA) module is composed of a convolutional LSTM (ConvLSTM) network [59], which is an extension of LSTM [21] and is responsible for modeling the spatial-temporal representations of the gait, and a multi-layer perceptron (MLP) used to aggregate the temporal features of $\hat{\mathbf{F}}$. ConvLSTM is expected to outperform LSTM-based approaches because it captures spatio-temporal correlations simultaneously. The spatial-temporal feature $\hat{f}_{\text{aggr}} \in \mathbb{R}^{128}$ can be extracted by the TFA module from $\hat{\mathbf{F}}$ as follows:

$$\hat{f}_{\text{aggr}} = \text{TFA}(\hat{\mathbf{F}}) \tag{2.18}$$

Positional Feature Concatenating

The Positional Feature Concatenating (PFC) module takes advantage of the walking speed information, which can be one of the invariant attributes for changes in distance measured from 3D LiDAR. As shown in Fig. 2.4, the PFC takes the average of all frames T in a re-weighted gait speed sequence $\mathbf{v}_{\text{gait}}$, and the extracted gait speed feature $f_{\text{speed}} \in \mathbb{R}$ is concatenated into $\hat{f}_{\text{aggr}}$ as follows:

$$\hat{f}^{\text{re}}_{\text{aggr}} = \text{Concat}(\hat{f}_{\text{aggr}}, f_{\text{speed}}) \tag{2.19}$$

Viewpoint-informed Feature Aggregating

The Viewpoint-informed Feature Aggregating (VFA) module is designed to extract more discriminative features by leveraging two viewpoints: the left-side view and back view. As illustrated in Fig. 2.5, this module aggregates the outputs of two networks, $f_{\text{left-side}}, f_{\text{back}} \in \mathbb{R}^N$, which are pre-trained from two different viewpoints as follows:

$$f_{\text{gait}} = \text{Avgpooling1D}(f_{\text{left-side}}, f_{\text{back}}), \tag{2.20}$$

where N donates the number of subjects used during training. The one-dimensional average pooling operator is adopted for effectively aggregating the two feature vectors, inspired by studies in 3D classification [25, 66].

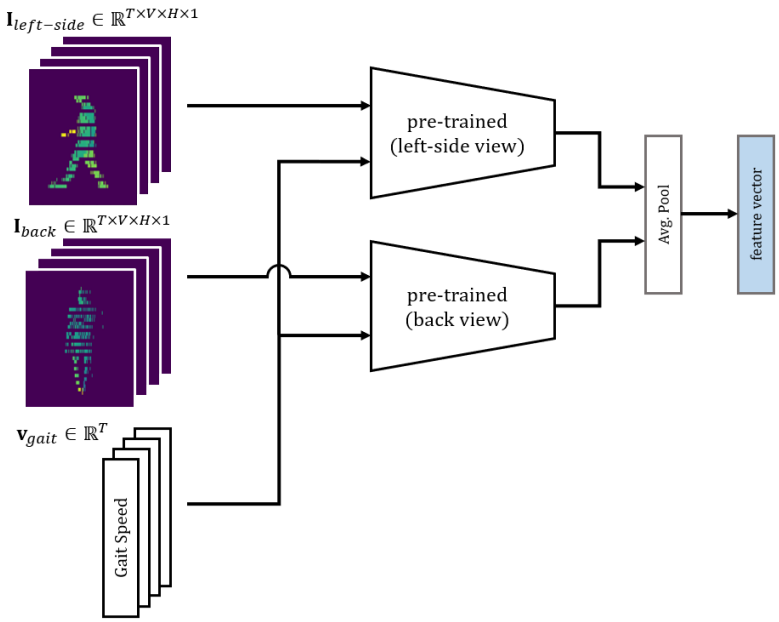


Figure 2.5: Detailed structure of VFA module, which aggregates the gait features of the recognition networks pre-trained from two different viewpoints: the left-side and back views.

2.5 Experiments

In this section, I describe the extensive experiments conducted on the point cloud gait datasets, which include variations in gait direction and measurement distance from the sensor. The results demonstrate the robustness of the proposed method to these changes. The dataset and training process are detailed first. Next, an ablation study is performed to verify the effectiveness of each component in the proposed model. Finally, the performance of the proposed method is compared against that of previous approaches.

2.5.1 Implementation Details

Four point cloud gait datasets with varying gait directions and measurement distances were used in this experiment. Each dataset contains data from 31 subjects, with each subject contributing 210 sequences. For training the networks, each dataset was divided into training, validation, and test sets, consisting of 140, 35, and 35 sequences, respectively. The first 16 subjects were grouped for training, while the remaining 15 subjects were used for testing.

During the training phase, all four training and validation sets from the first 16 subjects were utilized. The training data comprisedf $16\times4\times140 = 8960$ LiDAR gait sequences, while the validation data consisted of $16 \times 4 \times 35 = 2240$ LiDAR gait sequences. For optimization, Categorical Cross-Entropy was employed to train the network without incorporating metric learning. Furthermore, the

networks were trained using backpropagation with the RMSProp optimizer, a learning rate of 0.001, and a training batch size of 20 per subject. To prevent overfitting, I applied early stopping with a patience parameter of 20 [8, 84]. During the testing phase, the three training sets of the 15 subjects not used in the training phase were treated as gallery data. Each test set from the 15 subjects was then used as probe data, resulting in four test patterns. For inference, the network identified each subject using the nearest neighbor algorithm (Rank 1), which computed cosine similarity between the gallery and probe data.

2.5.2 Ablation Study

The quantitative results of the ablation study are presented in Tab. 2.1. The ablation study was conducted on the collected dataset to verify the effectiveness of each component by progressively adding the proposed modules individually. As shown in Tab. 2.1, identification accuracy was generally higher when the probe set included gait patterns, such as walking direction and measurement distance, that matched those in the gallery set, compared to when they did not. Comparing the results within the range of 11–13 m, I achieved a better performance by gradually adding the proposed modules. Especially, it is worth noting that the overall accuracy was greatly improved when applying the proposed VFA module. These results indicate that the proposed approach is effective and robust for recognizing gait features in sparse data by leveraging spatial representations of two different resolutions, the walking speed information, and two invariant viewpoints. In the results with the PFC module, accuracies were observed to decrease for all cases except for the highly sparse pedestrian data within the range of 11–13 m. One possible reason is that the walking speed information of pedestrians in the PFC module may not generalize well when learning gait features across the entire range from the sensor.

2.5.3 Comparison with Related Prior Methods

The quantitative results comparing the proposed network with previous work are presented in Tab. 2.1. In these results, the identification performance of the proposed network is compared with that of GEINet [61] and the previous method, LSTMNet [79]. GEINet, which uses a single GEI image as input to a convolutional neural network, is one of the most successful gait recognition methods and shares the same architecture as LGEI [6]. In contrast, LSTMNet is an LSTM-based network that identifies individuals using a depth image sequence acquired from a 3D LiDAR sensor. For both GEINet and LSTMNet, the left-side view gait image sequence, $\mathbf{I}_{\text{left-side}}$, was used as input. In the case of GEINet, $\mathbf{I}_{\text{left-side}}$ was transformed into a GEI image before being processed. Compared to GEINet and LSTMNet, the proposed network demonstrated significantly better performance when all components were applied. Notably, the proposed network showed considerable improvements in discriminative performance for individual recognition at long distances. The results of the quantitative evaluations confirm that the proposed model, which leverages the unique characteristics of 3D

Table 2.1: Averaged rank-1 accuracies on the collected 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 (proposed) → 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 (proposed) → 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 (proposed) → 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 (proposed) → 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		
GEINet [61] (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 [79] (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		

LiDAR sensors, enhances discriminative capability for gait recognition.

2.6 Conclusion

I propose a gait recognition method using a 3D LiDAR sensor to improve robustness against variations in measurement distance and walking direction. To enhance discrimination performance, I introduce a multi-view projection-based network that utilizes gait image sequences from two invariant views (left-side and back view) through an orthogonal projection strategy. To generalize gait features under variations in data sparsity, I propose a multi-scale spatial encoding and walking speed encoding approach. I also constructed a dataset for the gait recognition task using 3D LiDAR data, incorporating variations in range and gait direction. Finally, experiments conducted on the collected dataset demonstrate the effectiveness of the proposed method.

Chapter 3

Identification Modeling through Adaptive Learning

3.1 Introduction

In the previous chapter, I introduced a gait-based identification model using a 3D LiDAR sensor to minimize the impact of variations unrelated to gait features in two key aspects: viewing angles and measurement distances. These conditions are well-known as primary challenges typically encountered in LiDAR-related tasks. Specifically, I employed a two-scale spatial resolution approach to learn from varying point cloud densities projected onto 2D grids representing depth information. Moreover, the model leverages gait features from two invariant viewpoints (i.e., left-side and back views) across the gait sequence to enhance the consistency of walking dynamics, which cannot be captured by RGB cameras. The experimental results demonstrated superior performance compared to existing methods, highlighting the potential for applying 3D LiDAR-based gait recognition in such environments.

Despite these valuable results, however, there are still potential challenges to consider. First, the gait features from two viewpoints in this model are learned separately, which could result in issues with inference speed and optimization during training. In particular, the impact of self-occlusion on gait shapes varies depending on the emitting direction of the LiDAR sensor. Second, there is a need to develop a new dataset for evaluating identification performance. The dataset used in the experiments from the previous chapter includes simultaneous variations in measurement distance and walking direction, making it necessary to conduct separate performance evaluations for cross-view and cross-distance scenarios.

To address these issues, this chapter introduce an identification model that builds on a modified version of the model presented in the previous chapter, as illustrated in Fig. 3.1. Specifically, I

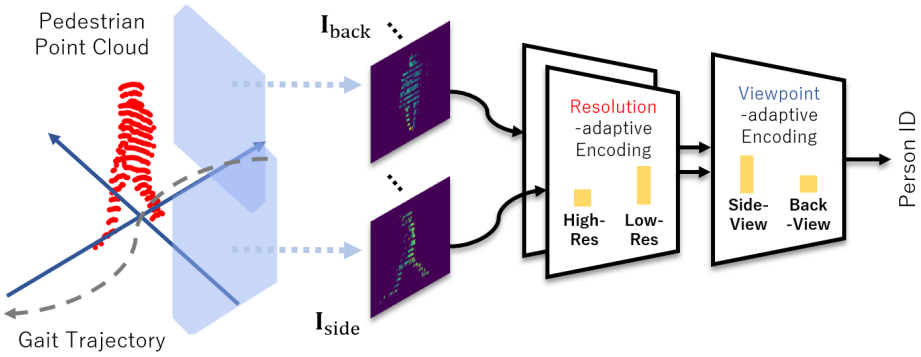


Figure 3.1: Overview of the proposed identification model using 3D LiDAR, which learns two viewpoint-invariant gait shapes in varying point cloud densities using an attention-based approach.

propose an attention-based block to fuse gait features from multiple resolutions and viewpoints. Unlike a typical self-attention mechanism that focuses on the interrelationships within a single input feature, this block compares the statistics of two different features, biasing towards the more informative one. I have extended this version based on the following three aspects: 1) Rather than using pooling approaches [25, 66], I designed a novel attention block to fuse the two gait features more effectively for both invariant viewpoint and spatial resolution in an end-to-end manner. 2) I conducted an in-depth ablation study to evaluate the proposed modules, as well as the effectiveness of orthogonal projection and depth information. 3) I compared the performance with existing methods on the new dataset, which consists of combinations of cross-views and cross-distances, to achieve deeper insights than previous study.

The contributions of this chapter are as follows:

1. I proposed a framework for gait recognition using 3D LiDAR sensor. This model is robust to changes in viewing angles and measurement distances.

2. Performance of the proposed method is enhanced in three aspects, namely, point cloud projection, gait direction transformation, and recognition network, to learn viewpoint- and point cloud density-independent gait features with contextual depth information.

3. Extensive experiments, including variances of both viewing angles and measurement distances, conducted on the dataset revealed that the proposed method surpassed the recognition accuracy of prior studies.

3.2 Related work

3.2.1 Attention Mechanism for Convolutional Neural Networks

The attention mechanism has been used in numerous tasks because of its ability to bias the allocation of available processing resources toward the most informative components of an input signal [73]. Several studies have demonstrated its applicability to computer vision. For example, Wang *et al.* [74] proposed a nonlocal operation that captures long-range dependencies in images or videos and can be plugged into CNN-based architectures. Hu *et al.* [22] enhanced the representation power of CNNs by focusing on channel relationships using a gating mechanism, whereas Woo *et al.* [75] inferred attention maps related to both the channel and spatial features. In this study, a novel attention block was designed by considering inter-channel dependencies to effectively fuse two gait features extracted from convolutional encoders.

3.3 Method

In this section, I demonstrate a pipeline for the proposed gait recognition method that consists of three steps: gait direction transformation, depth image generation, and recognition network. Specifically, the gait direction transformation and depth image generation processes are slightly refined versions of those in the previous chapter. To enhance immunity to changes in the walking directions of subjects, I used LiDAR characteristics, such as gait shapes that are invariant to viewing angles and are not captured by general RGB cameras.

3.3.1 Gait Direction Transformation

I first describe the gait direction transformation (GDT) process for estimating the walking directions of point cloud sequences. The gait directions are transformed into constant directions to extract the two viewpoint-invariant gait features. For example, when generating gait images from a left-side view, subject point sets are aligned with the $-y'$-axis in the new $x'y'$-plane of Cartesian coordinates, to project these side gait shapes from the $y'z'$-plane, as depicted in Fig. 3.2. Let $\mathbf{P}_t = \{\mathbf{p}_{t,1}, \mathbf{p}_{t,2}, \ldots, \mathbf{p}_{t,N}\}$ denote the point set of a subject at time step t, obtained using either background subtraction or object tracking techniques. Here, N denotes the number of points for time step t, which may vary across different frames. In addition, n represents an arbitrary single point among N points. Each point $p_{t,n} \in \mathbb{R}^3$ is represented in Cartesian coordinates $(p_{t,n,x}, p_{t,n,y}, p_{t,n,z})$, where the z-coordinate is a vertical directional value at the corresponding points. Given an extracted point cloud of the subject, the center of mass $\mathbf{c}_t = (c_{t,x}, c_{t,y}, c_{t,z})$ for the time step t is defined as follows:

$$\mathbf{c}_t = \frac{1}{N} \sum_{n=1}^{N} \mathbf{p}_{t,n}, \tag{3.1}$$

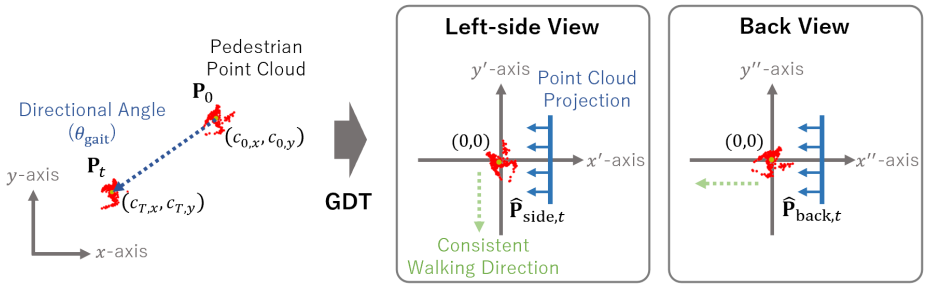


Figure 3.2: Overview of the gait direction transformation (GDT) process that generates two invariant gait shapes from pedestrian point cloud sequences.

where $c_{t,z}$ is set to zero because only the walking direction on the xy-plane is considered. Subsequently, the directional angle θ_{gait} for a given point cloud sequence $\mathbf{P}$ in the xy-plane can be calculated from its starting and ending central points. This approach avoids instability of the central points caused by a variable number of points in the gait sequence. All walks over an entire frame can be approximated as a straight line:

$$\theta_{\text{gait}} = \arctan(c_{T,y} - c_{0,y}, c_{T,x} - c_{0,x}), \tag{3.2}$$

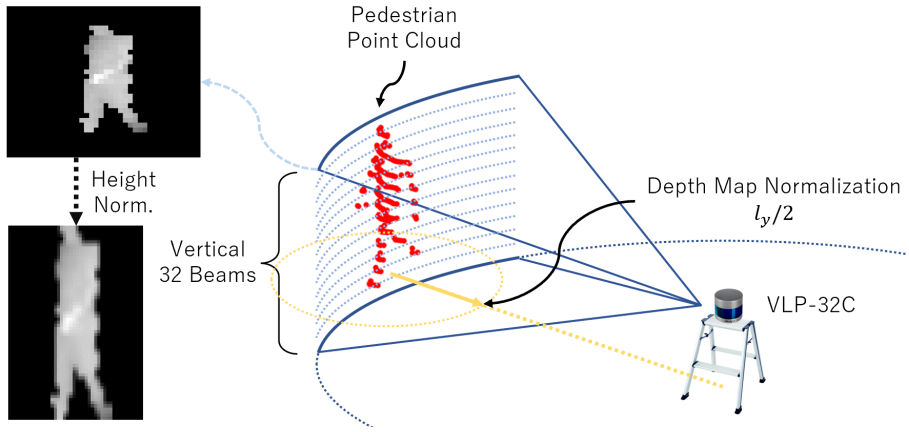
where $\arctan(\cdot,\cdot)$ calculates the angle between a pedestrian and a positive x-axis in the Euclidean plane. Unlike the GDT process in the previous version, where the gait angle θ_{gait} for each frame t is calculated using the immediately preceding and succeeding frames, this updated process calculates θ_{gait} based on T and 0 across the entire gait sequence. This adjustment aims to reduce errors in sparse pedestrian point clouds measured at long ranges. Given a directional angle θ_{gait}, the transformed point cloud sequence $\hat{\mathbf{P}} = \{\hat{\mathbf{P}}_0, \hat{\mathbf{P}}_1, \ldots, \hat{\mathbf{P}}_T\}$ is obtained by rotating the original gait sequence $\mathbf{P}$ around the central points using the z-axis in the case of a left-side view:

$$\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_{\text{gait}} - \frac{\pi}{2}) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t), \tag{3.3}$$

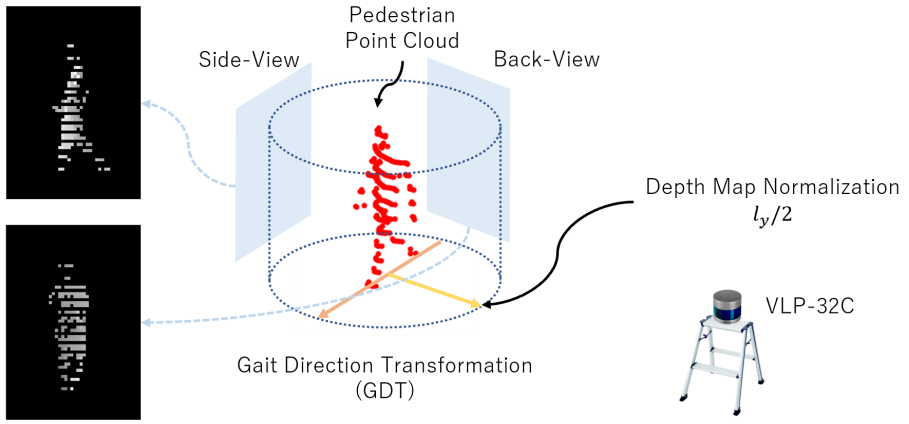
where $\mathbf{R}_z$ represents a rotation matrix around the z-axis. The case of generating back-view gait images follows the above procedure, except that rotation matrix $\mathbf{R}_z(-\theta_{\text{gait}} - \pi/2)$ and the $y'z'$-plane are replaced with $\mathbf{R}_z(-\theta_{\text{gait}} - \pi)$ and the $y''z''$-plane, respectively, as shown in Fig. 3.2. The point clouds from the left-side and back views are standardized to be represented as depth images in the same coordinate direction by varying the values of the rotation matrix $\mathbf{R}_z(\cdot)$.

3.3.2 Depth Image Generation

Input data are generated from the transformed pedestrian point cloud $\hat{\mathbf{P}}$ to feed a subsequent recognition network: depth image sequences of the left-side and back views. In this study, the gait speed



(a) Spherical Projection



(b) Orthographic Projection

Figure 3.3: Comparison between the prior spherical and proposed orthographic projection approaches, representing gait shapes with depth information.

sequence is not used for identification due to challenges in generalizing it during network training. Additionally, unlike the previous version, a constant is introduced to normalize the height of pedestrians in the gait videos. This approach represents point clouds as gait images using orthographic projection, which allows for a more intuitive representation of walking shapes and texture information from fixed viewing angles. This projection manner differs from the approach used in [79], in which point clouds are assigned to an angular grid using spherical projection, as displayed in Fig. 3.3. The proposed gait images are rendered from the subject point cloud $\hat{\mathbf{P}}_t \in \mathbb{R}^{N\times 3}$ heading along the $-y$-axis and $-x$-axis, which correspond to the left-side and back views, respectively. The gait image $\mathbf{i}_t$ of each time step t is determined as $V(= l_z/R_z) \times H(= l_y/R_y)$ grid, where l_z, l_y, R_z, and R_y are a height for the z-axis, width for the y-axis, vertical resolution, and horizontal resolution, respectively, for the generated images. Compared with the prior approach in [79], which bijectively maps the pedestrian point cloud to a 2D spherical grid through one-to-one correspondence, the proposed method assigns this point cloud to a physical space divided into pixel-level regions. When more than one point exists in the same pixel, the largest x-coordinate value, representing the nearest point based on the direction $-x$-axis in which gait shapes are observed, is the depth value for that corresponding pixel. If no points exist within a pixel, its depth value is set to 0. This method for determining depth values is similar to the Z-buffer algorithm, which compares the depths of surfaces at each pixel position on the projection plane, except that it uses the smallest depth values. For a clear distinction between a pedestrian and the background, a constant $l_y/2$ is added to all pixel values where one or more points exist. Here, the pixel position $i_{v,h}$ in the proposed depth images for an arbitrary point $\hat{\mathbf{p}}_{t,n}$ is defined as follows:

$$v = \left\lfloor \frac{1}{R_z} \cdot (\hat{p}_{t,n,z} - \min_{n\in\{1,\dots,N\}} (\hat{p}_{t=0,n,z}) + l_{z-\text{const}}) \right\rfloor, \tag{3.4}$$

$$h = \left\lfloor \frac{1}{R_y} \cdot (\hat{p}_{t,n,y} + \frac{l_y}{2}) \right\rfloor, \tag{3.5}$$

where the criterion for the vertical axis in the generated gait images is determined by considering the lowest z-coordinate value of the point cloud sequence $\hat{p}_{t=0,n,z}$ at time step $t = 0$, which represents the floor in walking situations, with an additional constant $l_{z-\text{const}}$, to standardize gait shapes projected onto the image sequences. Compared with typical gait recognition tasks that involve resizing a pedestrian segmented from RGB images to a standard height via linear interpolation, the proposed approach directly projects a subject's point cloud onto the image. Therefore, the proposed gait images require limited pre-processing than general RGB images and provide richer size-related information. Consequently, the gait image sequence $\mathbf{I} = (\mathbf{i}_1, ..., \mathbf{i}_T) \in \mathbb{R}^{T\times V\times H}$ for T frames is obtained. We used depth images projected from two fixed viewpoints for the subsequent recognition network: gait image sequences of left-side view $\mathbf{I}_{\text{side}}$ and back view $\mathbf{I}_{\text{back}}$. Although $\mathbf{I}_{\text{side}}$ provides the richest dynamic gait information, $\mathbf{I}_{\text{back}}$ is more practical than other viewing angles because

people tend to walk away from visual sensors and prefer not to show their faces to visual sensors. The proposed depth images can represent gait-related features more effectively than the silhouettes and RGB images typically used in gait recognition tasks because these images convey geometric information more clearly than other images. The depth values for each gait image sequence are normalized by dividing them by a constant $l_y/2$. This normalization step can scale the depth values within a specific range and facilitate the training process of the proposed network. In this study, l_z, l_y, R_z, R_y, V, H, $l_{z-\text{const}}$, and T are set to 2.6 m, 1.8 m, 0.04 m, 0.04 m, 64, 44, 0.4 m, and 15, respectively. Here, V and H are selected based on previous gait recognition studies and are used consistently throughout the experiments.

3.3.3 Recognition Network

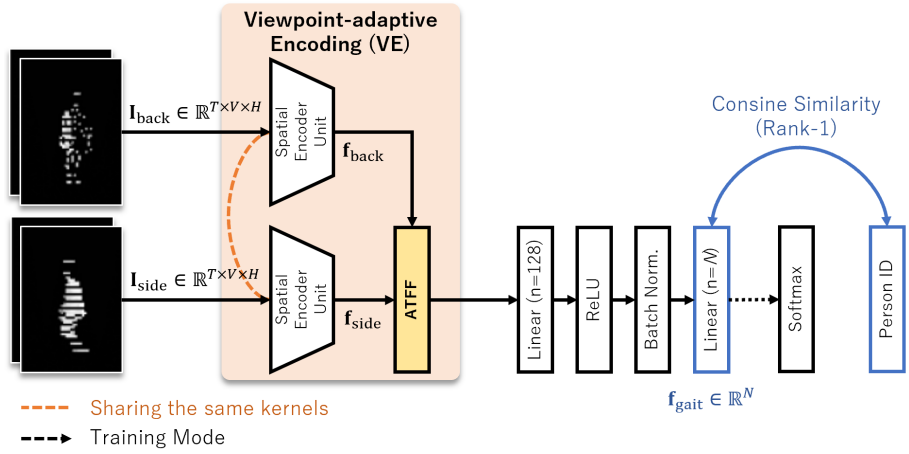
In this section, I describe a recognition network for learning the discriminative gait features from the aforementioned inputs. The architecture in Fig. 3.4a consists of a viewpoint-adaptive encoding (VE), which includes two spatial encoder units and one attention-based two-feature fusing (ATFF) block, along with additional layers. Each spatial encoder unit in the VE module is formed of two components, namely, resolution-adaptive encoding (RE) and temporal encoding (TE), as displayed in Fig. 3.4b. In particular, the ATFF block is inserted into the ends of the VE and RE to flexibly aggregate the two features under various confounding conditions.

Viewpoint-Adaptive Encoding

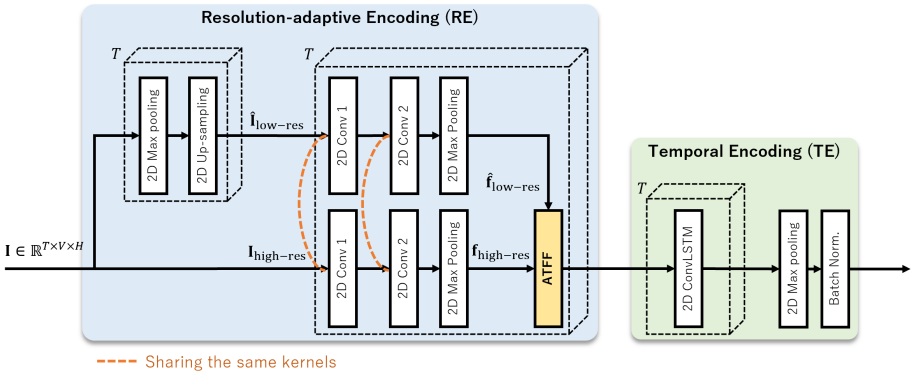
In the viewpoint-adaptive encoding (VE) module, two-feature maps are fused from different viewpoints: the left-side and back views to obtain the discriminative gait feature. Specifically, the gait features $\mathbf{f}_{\text{side}}$ and $\mathbf{f}_{\text{back}}$ are extracted from the same spatial encoder unit and combined into one feature vector through the ATFF block. Unlike the previous version of this study [1], which aggregates outputs from two units pre-trained for different viewpoints, this feature fusion is achieved in an end-to-end manner. Subsequently, the final linear layer in the proposed final network is used as a gait feature vector $\mathbf{f}_{\text{gait}} \in \mathbb{R}^N$, where N is the number of trained subjects. During training, I adopt cross-entropy loss, which is common in classification tasks, and calculate the gap between a predictive distribution and the corresponding ground-truth distribution. In contrast, during testing, I use the nearest neighbor algorithm (i.e., Rank-1 accuracy) to compute the cosine similarity between galleries and probes for subsequent evaluations.

Resolution-Adaptive Encoding

The high-resolution images represent fine-grained gait patterns. However, when the input data are captured at a long distance, these images do not have the ability to recognize detailed spatial features because the human shape is generally sparse, as displayed in Fig. 3.5. Gait-related spatial features



(a) Overall recognition network



(b) Spatial encoder unit

Figure 3.4: Architecture of (a) overall recognition network and (b) spatial encoder unit. (a) The overall recognition network consists of a viewpoint-adaptive encoding (VE) module, two fully-connected layers, a ReLU activation function, and batch normalization. Specifically, the VE module includes two spatial encoder units to process gait image sequences of both the left-side and back views. (b) The spatial encoder unit is equipped with a resolution-adaptive encoding (VE) module and a temporal encoding (TE) module. This unit extracts the gait feature for a single viewpoint.

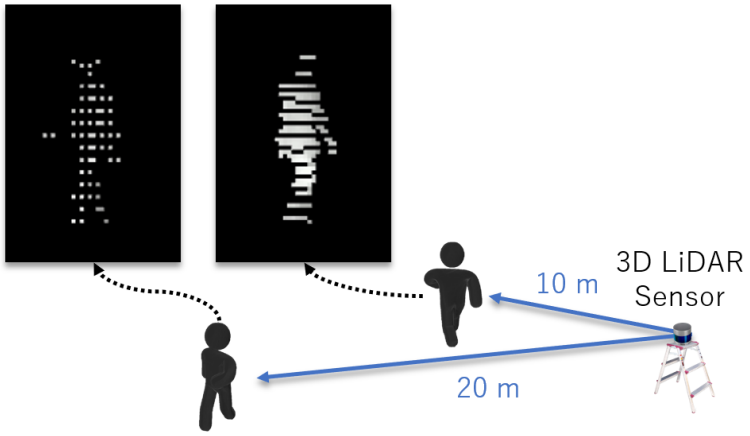


Figure 3.5: Examples with various measurement distances, with different sparsity of proposed depth images.

learned from dense data acquired at short distances may not generalize to long distances, as the point clouds captured from LiDAR sensors typically have non-uniform densities at various distances. To alleviate this problem, I designed an RE module that leverages not only the original resolution but also the low resolution. This method is robust to sparse data and exhibits an enhanced recognition of coarse-grained patterns. Furthermore, this module combines the two-scale features extracted from the two resolutions. First, the double-reduced-resolution image sequence $\mathbf{I}_{\text{low-res}} \in \mathbb{R}^{T\times V/2\times H/2}$ is obtained by feeding a gait image sequence $\mathbf{I} \in \mathbb{R}^{T\times V\times H}$ into a max pooling layer as follows:

$$\mathbf{I}_{\text{low-res}} = \text{Maxpool2D}(\mathbf{I}) \tag{3.6}$$

$\mathbf{I}_{\text{low-res}}$ is up-sampled to match the height and width dimensions of the original image $\mathbf{I}$ as follows:

$$\hat{\mathbf{I}}_{\text{low-res}} = \text{Upsampling2D}(\mathbf{I}_{\text{low-res}}) \tag{3.7}$$

Subsequently, two gait image sequences are fed into the CNN-based extractor. This extractor consists of two 2D convolutional layers and one 2D max-pooling layer to extract two spatial feature maps with two different resolutions from a single viewpoint. In this extractor, the kernels are shared across both two resolutions and two viewpoints to reduce the computational cost and to learn filters with the same weights for the subsequent ATFF block in the RE module. The detailed configurations of each layer are listed in Table 3.1.

Table 3.1: Layer configuration for the spatial encoder unit.

Layer	Input/Output Channels	Kernel Size	Stride/Padding	Activation
2D Conv 1	1/32	5×5	1/0	ReLU
2D Conv 2	32/32	5×5	1/0	ReLU
2D Max Pooling	-	2×2	2/0	-
2D ConvLSTM	32/64	3×3	1/0	-

Attention-Based Two Features Fusing

As displayed in Fig. 3.6, the ATFF blocks can adaptively recalibrate and aggregate two-feature maps under changing conditions, especially in terms of resolutions and viewpoints. For instance, high-resolution images represent distinct spatial features at short distances from LiDAR sensors, whereas low-resolution images may be more effective at long distances, as displayed in Fig. 3.5. Additionally, a more optimal viewpoint for capturing gait-related features could exist because self-occlusions depend on the emitting angles of the lasers. Inspired by [22], I designed a novel attention-based block that fuses two 2D feature maps that represent distinct characteristics, biasing more useful weights under varying scenarios. Unlike the typical self-attention mechanism that explores the inter-relationships within a single input, this block operates channel-wise comparisons between two-feature maps. Thus, in this approach, the scores for both inputs are compared and fused into a single feature. Given that inputs fed into the ATFF module are denoted as $\mathbf{f}_1 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$ and $\mathbf{f}_2 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$, the global spatial information is compressed using global average pooling in each channel to fully exploit the contextual information. Here, C_{attn}, H_{attn}, and W_{attn} represent the channel, height, and width of these 2D spatial features, respectively. Compared with the original structure [22], my strategy incorporates an additional temporal context T into the global average pooling calculation to capture the gait-related consistency in sequences. Formally, the vector $\mathbf{z}_1 \in \mathbb{R}^{c_{\text{attn}}}$ is calculated with the spatio–temporal dimensions $T\times H_{\text{attn}}\times W_{\text{attn}}$ of $\mathbf{f}_1$ for each channel by the following equation:

$$z_{1,c_{\text{attn}}} = \frac{1}{T\times H_{\text{attn}}\times W_{\text{attn}}}\sum_{i=1}^{T}\sum_{j=1}^{H_{\text{attn}}}\sum_{k=1}^{W_{\text{attn}}} f_{1,c_{\text{attn}}}(i,j,k). \tag{3.8}$$

Subsequently, this operation is followed with a second operation that fully captures channel-wise dependencies and expresses probabilistic values to determine which of the two features $\mathbf{f}_1$ and $\mathbf{f}_2$ is more critical. First, the statistical vector $\mathbf{s}_1 \in \mathbb{R}^{C_{\text{attn}}}$ of $\mathbf{f}_1$ is obtained by forming a bottleneck using a dimensionality reduction layer with reduction ratio r and sigmoid activation. Furthermore, another statistic $\mathbf{s}_2 \in \mathbb{R}^{C_{\text{attn}}}$ of $\mathbf{f}_2$ is calculated with $\mathbf{s}_1$ as follows:

$$\mathbf{s}_2 = \mathbf{1} - \mathbf{s}_1 = \mathbf{1} - \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z}_1)), \tag{3.9}$$

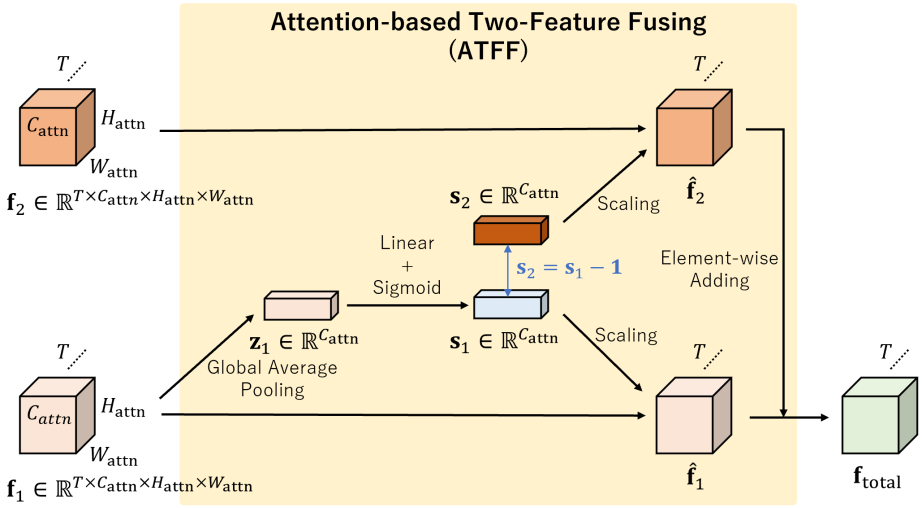


Figure 3.6: Structure of the attention-based two features fusing (ATFF) block, which takes two different feature maps as input and recalibrates their scores to fuse them into a single feature.

where $\sigma(\cdot)$ and $\delta(\cdot)$ refer to the sigmoid and ReLU functions with $\mathbf{W}_1 \in \mathbb{R}^{\frac{C_{\text{attn}}}{r}\times C_{\text{attn}}}$ and $\mathbf{W}_2 \in \mathbb{R}^{C_{\text{attn}}\times\frac{C_{\text{attn}}}{r}}$. Here, the sum of $s_{1,c_{\text{attn}}}$ and $s_{2,c_{\text{attn}}}$ for each channel is equal to 1 so that each statistic is standardized. This subtraction operation is designed based on the insight that the channel-wise scores of one input are determined simultaneously when the scores of another input are computed. Finally, the final output $\mathbf{f}_{\text{total}} \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$ of this ATFF block can be obtained by adding two outputs $\hat{\mathbf{f}}_1$ and $\hat{\mathbf{f}}_2 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$, which are rescaled from the inputs $\mathbf{f}_1$ and $\mathbf{f}_2$ using the activation $\mathbf{s}_1$ and $\mathbf{s}_2$, respectively:

$$\mathbf{f}_{\text{total}} = \hat{\mathbf{f}}_1 \oplus \hat{\mathbf{f}}_2 = (\mathbf{s}_1 * \mathbf{f}_1) \oplus (\mathbf{s}_2 * \mathbf{f}_2), \tag{3.10}$$

where $\oplus$ and $*$ represent element-wise adding and channel-wise multiplication between a scalar $s_{c_{\text{attn}}}$ and a feature map $f_{c_{\text{attn}}}$, respectively. These ATFF blocks are inserted at the endpoints of the RE and VE modules in the proposed network. In the VE module, the T dimension of the input feature is set to 1.

Temporal Encoding

The TE block aggregates the temporal information of the feature maps extracted from the previous RE module. Specifically, this module consists of a single 2D convolutional LSTM (ConvLSTM) layer [59], which is an extension of LSTMs [21] and is used for modeling the spatio–temporal representation of gait features, equipped with 2D max pooling and batch normalization layers [23]. The hyper-parameters of this layer are listed in Table 3.1. Compared with [79], ConvLSTM can outperform 1D-LSTM layers because it captures spatio–temporal correlations simultaneously. In Addition, the

kernels in this TE module are shared across two viewpoints for the subsequent ATFF block in the VE module.

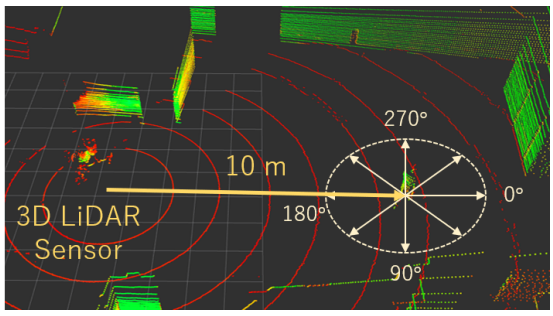
3.4 Datasets

In this section, I demonstrate the collected dataset to evaluating the effectiveness of the proposed model. To verify the robustness to changes in viewing angles and distances measured from a sensor, I collected a gait database using a 32-beam LiDAR scanner, Velodyne VLP-32C, which creates 3D range images with a horizontal 360° field of view using 32 lasers for vertical resolution. Compared to the HDL-32C used in the previous chapter, its vertical lasers are more concentrated toward the center, enabling it to capture objects more densely at the same measurement range. This data consists of gait sequences containing 30 subjects in a 3D point cloud format. Furthermore, in this dataset, the sampling rate was set to 10 Hz and it had 15 frames. During data collection, the LiDAR sensor was mounted on a tripod at a height of 1.2 m. I then requested each subject to walk as usual along four straight lines that evenly divided a circle, located at distances of 10 and 20 m away from the sensor, as displayed in Fig. 3.7. I obtained gait data for each subject under eight views (0°, 45°, 90°, 135°, 180°, 225°, 280°, 315°) and two distances (10 and 20 m) per subject. For the subsequent experiments, the viewing angles are determined by the pedestrian's walking direction relative to the sensor-pedestrian vector, as shown in Fig. 3.5. Compared with other well-known gait datasets [55,83] commonly used in gait recognition challenges, this combination not only includes cross-view but also cross-distance conditions. In this dataset, 126-point cloud sequences were obtained for each subject under a single condition. Therefore, the gait dataset contains $30 \times 8 \times 2 \times 126 = 60{,}480$ sequences. During training and evaluation, I only used the pedestrian point cloud sequences, which were extracted through background subtraction processing.

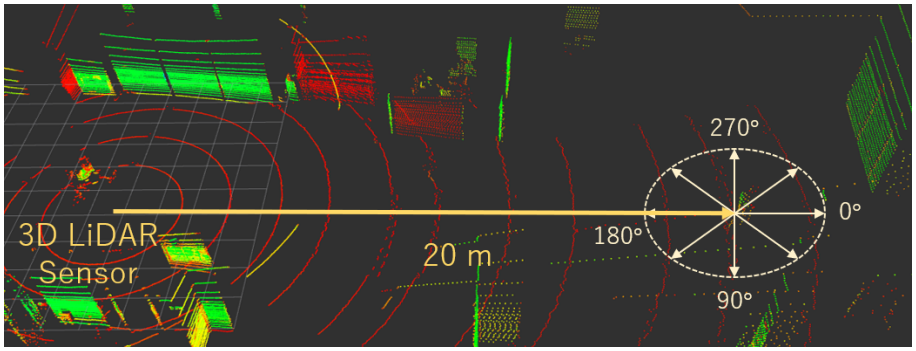
3.5 Experiments

3.5.1 Implementation Details

I conducted experiments based on the collected dataset. Essentially, the first 20 subjects were used for training, and the remaining 10 subjects were used for testing with no overlap. Thus, the training set contains $20 \times 8 \times 2 \times 126 = 40{,}320$ sequences for all experimental settings. In addition, I standardized the input to a set of aligned images with a size of 64×44 in all recognition networks to ensure a fair comparison with prior studies, as displayed in Fig. 3.8. For optimization, I used the cross-entropy loss to train the networks. I adopted ADAM [26] optimizer for optimization. The details for batch size, learning rate, and training iterations in all the experimental settings were 42, 1e-4, and 48k, respectively. The code for all experiments was implemented using Python in PyTorch 1.12, and performed on a single NVIDIA GeForce RTX 3090 GPU (approximately 10 hours to train).

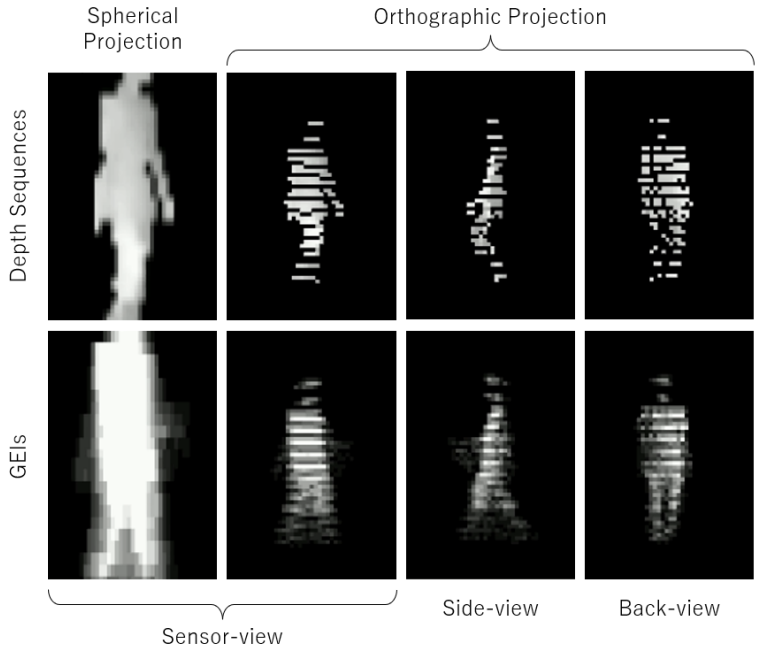


(a) Measurement distance: 10 m

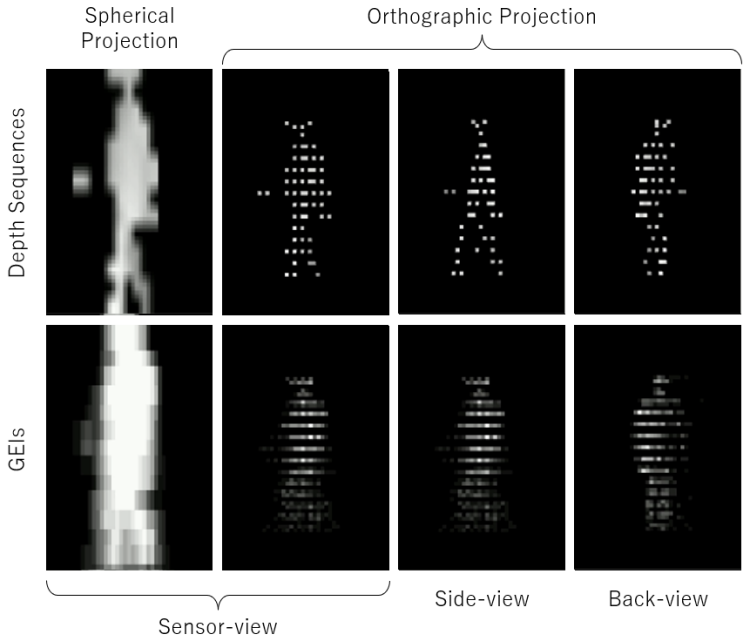


(b) Measurement distance: 20 m

Figure 3.7: Data acquisition environment with two distances measured from a VLP-32C, which is visualized in a 3D point cloud format.



(a) Measurement distance: 10 m



(b) Measurement distance: 20 m

Figure 3.8: Visualization of input data for feeding into recognition networks with two distances measured from a VLP-32C, which are generated through a combination of point cloud projection, viewpoint, and modality.

During testing, I used the penultimate feature to match the 10 subjects that were not used in the training.

3.5.2 Method Comparison

To evaluate the contribution of each component to the overall performance, I investigated the effectiveness of the proposed method by examining it from three changeable perspectives: viewpoint, point cloud projection, and recognition network. In the GDT, I investigated the extent to which viewing angle-independent gait features, which cannot be directly obtained from RGB cameras, contribute to individual recognition. Furthermore, I evaluated the proposed point cloud projection for input and compared it with the existing method (i.e., spherical projection) used in [79]. When applying the spherical projection, I normalized gait images to a size of 64 $\times$ 44 based on the height of pedestrians segmented from the bijective 2D grids with bilinear interpolation, in which pre-processing is commonly used in camera-based gait recognition tasks. In the recognition network part, I compared the performance of the proposed network with three prior approaches [6, 61, 79]. Among these approaches, the LGEI-based technique [6] is the first gait analysis study using LiDAR sensors. This network feeds the GEIs of a gait sequence into the CNN layers and performs person re-identification. On the other hand, [61], called GEINet, is the most representative network that uses GEIs in the gait recognition field, which structure is similar to that of [6]. In [79], gait image sequences with depth information are input into CNN and LSTM layers to classify individuals. I compared the proposed method with both Network 1 and 2 in [79], where the second architecture is a modified version of the first method by supplementing a subtraction operation at the front of the LSTMs.

3.5.3 Main Results

Table 3.2 presents a comparison between the proposed method and the prior approaches. The results for all networks were obtained through experiments. In this experiment, I averaged all results across seven views, excluding identical views. Furthermore, the robustness of the proposed method was evaluated in various point cloud densities by varying the measurement distances between the galleries and probes. This experimental setting is more challenging than the typical cross-view conditions. In this experiment, the gallery and probe contained 10 $\times$ 1 $\times$ 7 $\times$ 42 = 2,940, and 10 $\times$ 1 $\times$ 1 $\times$ 84 = 840 sequences for each condition, respectively. In the proposed network with a single viewpoint, I used only one spatial encoder unit in the VE module without the ATFF block.

First, in Table 3.2, I observed that networks using the proposed point cloud projection (Ortho.) achieved higher average accuracies than the prior approach (Spher.). A possible reason for this phenomenon is that the previous method requires linear interpolation processing to adjust the size of pedestrians to the same height, which may exclude individuals' unique spatial gait features, such as stride or height of the body. Although the generated images are sparse at long distances,

in the proposed projection, real-size gait shapes are used without pre-processing, such as linear interpolation algorithms. I achieved improved recognition performance by using both the GDT and the proposed projection (Ortho.) except for Network 1 [79]. A potential reason is that the lateral gait shapes can capture dynamic gait information more clearly, such as swinging motions of the arms or legs, compared with the original view from the sensors. On the other hand, when applying the prior projection (Spher.) to the GDT process, I observed a performance decline in the networks that use depth image sequences [79] [1]. This result indicates that the spherical projection method may lead to increased distortion of the gait shapes when rotating pedestrian point clouds around the z-axis, compared with the proposed projection. Furthermore, self-occlusions with LiDAR scanning can be one of the potential causes of the performance decline in Network 1 [79] because they negatively affect the depth values in gait images processed with the GDT. In the proposed network and the previous version [1], the average accuracies of the back view exceed those of the side view. Based on this result, the gait shapes captured from a back view, especially when depth information is included, could be critical discriminative cues for recognition. In addition, I observed that the fixed viewpoint strategy positively affected GEI-based networks [5, 61] in both the proposed and prior projection methods. This result indicates that GEIs captured from the side-view show more significant changes in the overall gait shapes than those from the original view. Finally, when comparing the accuracies across all combinations, the proposed complete method achieved the highest discriminative capability by using two viewpoints.

Table 3.3 presents the performance results for the recognition networks under the sole cross-view condition. This experimental setting is identical to that in Table 3.3, except that the measurement distances for both the gallery and probe are the same. Each average value in Table 3.3 represents a combination of all eight cross-views and two distances. I observed that the results in Table 3.3 are generally higher than those in Table 3.2 because this experiment evaluated only the robustness with changes in the walking direction. In particular, among the accuracy results for all combinations, the proposed method using two viewpoints achieved the second-highest accuracy, followed by the previous model [1]. From these results, I observed that while the pooling method that combines two gait features from independently trained units may achieve optimal performance under a cross-view condition, the proposed attention method is more effective in scenarios where the point cloud density changes.

3.5.4 Ablation Study

In this section, I report the ablation experiments on the collected dataset used in the comparison experiment to evaluate the effectiveness of the elements proposed in the proposed recognition network.

Table 3.2: Comparison with prior studies on the collected dataset under two conditions (%)

Distances		Networks	Modalities	Projections	Viewpoints			Gallery All 8 Views (0 – 315°)								
Gallery	Probe				Sensor-view	Side-view	Back-view	0°	45°	90°	135°	180°	225°	270°	315°	Mean
10 m	20 m	Benedek *et al.* [6]	GEI	Spher.	✓			27.3	32.0	42.5	28.6	26.1	33.1	36.9	29.6	32.0
						✓		41.6	51.8	46.4	25.4	40.6	42.7	51.1	41.9	42.7
				Ortho.	✓			58.3	57.4	48.0	55.8	55.0	43.1	46.6	49.2	51.7
						✓		52.7	58.3	63.3	65.0	64.6	55.0	50.2	56.8	58.2
		Shiraga *et al.* [61]	GEI	Spher.	✓			21.4	29.4	43.5	34.9	22.5	25.4	41.8	24.8	30.5
						✓		24.2	37.0	31.2	41.3	23.8	34.8	35.4	30.2	32.2
				Ortho.	✓			44.4	44.4	32.1	41.7	35.6	36.6	33.5	41.2	38.7
						✓		52.9	50.5	65.6	62.4	53.2	47.9	59.2	47.3	54.9
		Yamada *et al.* (Network 1) [79]	Depth Seq.	Spher.	✓			21.6	22.1	29.8	29.3	24.3	30.4	33.6	24.1	26.9
						✓		11.6	20.6	8.5	12.9	18.7	11.3	22.6	15.5	15.2
				Ortho.	✓			54.2	49.9	49.8	60.1	51.3	48.3	53.0	57.4	53.0
						✓		50.1	53.0	53.1	49.9	49.3	49.4	45.2	51.1	50.1
		Yamada *et al.* (Network 2) [79]	Depth Seq.	Spher.	✓			17.7	36.8	32.1	37.6	23.0	33.7	29.3	29.3	30.0
						✓		12.3	13.1	17.1	15.6	10.7	9.8	10.0	16.0	13.1
				Ortho.	✓			37.3	47.1	32.5	51.1	44.5	41.3	33.2	49.6	42.1
						✓		51.8	50.5	58.8	53.1	54.1	51.4	53.2	45.4	52.3
		Ahn *et al.* [1]	Depth Seq.	Spher.		✓	✓	52.3	67.9	55.6	69.4	66.9	62.1	56.2	54.9	60.7
				Ortho.		✓	✓	73.8	77.4	81.4	83.3	81.8	74.5	**89.1**	74.3	79.5
		Proposed	Depth Seq.	Spher.	✓			29.8	39.9	46.2	38.3	42.9	36.2	43.9	30.6	38.5
						✓		19.4	31.4	44.2	46.7	22.3	27.7	11.0	30.6	29.2
							✓	54.6	50.0	35.1	55.1	57.7	53.8	32.9	41.8	47.6
						✓	✓	50.8	51.8	52.6	60.0	60.7	59.3	48.6	46.4	53.8
				Ortho.	✓			64.5	73.3	65.6	72.5	68.2	73.9	63.6	71.0	69.1
						✓		62.3	64.6	80.2	73.0	74.1	69.4	80.8	64.5	71.1
							✓	**77.5**	72.7	69.1	78.5	81.7	78.7	71.4	68.9	74.8
						✓	✓	69.3	**82.9**	**90.0**	**85.4**	**84.9**	**82.4**	80.4	**78.2**	**81.7**
20 m	10 m	Benedek *et al.* [5]	GEI	Spher.	✓			26.2	30.1	19.2	31.7	22.0	28.3	32.5	29.6	27.5
						✓		50.5	47.7	41.4	36.0	40.5	37.5	40.8	50.4	43.1
				Ortho.	✓			57.4	53.5	52.0	52.0	56.4	54.9	50.4	49.5	53.3
						✓		52.1	53.5	56.4	58.8	58.6	58.5	57.0	52.4	55.9
		Shiraga *et al.* [61]	GEI	Spher.	✓			28.3	32.7	26.2	39.3	26.2	32.1	20.4	42.9	31.0
						✓		41.4	39.3	40.6	47.6	30.0	32.0	28.8	33.5	36.7
				Ortho.	✓			28.0	43.5	28.6	40.4	31.4	45.5	47.4	39.6	38.0
						✓		52.9	60.8	57.3	68.9	67.1	72.5	52.9	58.1	61.3
		Yamada *et al.* (Network 1) [79]	Depth Seq.	Spher.	✓			29.3	33.0	24.1	35.4	29.4	32.3	28.7	31.4	30.4
						✓		17.7	17.6	22.6	21.4	18.3	18.2	14.1	8.6	17.3
				Ortho.	✓			52.6	64.3	49.3	53.7	53.5	65.1	46.0	54.3	54.8
						✓		49.5	54.1	41.4	52.1	56.4	53.9	45.7	50.7	50.5
		Yamada *et al.* (Network 2) [79]	Depth Seq.	Spher.	✓			30.5	43.8	40.4	37.5	22.9	38.8	41.2	39.4	36.8
						✓		10.1	17.5	13.2	15.8	12.9	13.3	17.3	13.7	14.2
				Ortho.	✓			45.6	58.6	49.9	48.6	38.5	59.3	52.5	54.3	50.9
						✓		53.0	60.5	52.6	56.7	50.0	62.9	51.6	55.6	55.4
		Ahn *et al.* [1]	Depth Seq.	Spher.		✓	✓	58.8	62.4	65.8	74.5	72.0	69.6	72.5	69.3	68.1
				Ortho.		✓	✓	80.7	81.0	70.1	71.7	84.2	78.3	**78.6**	85.5	78.8
		Proposed	Depth Seq.	Spher.	✓			39.6	43.9	35.4	52.1	47.1	39.8	43.6	49.9	43.9
						✓		25.5	48.5	34.1	29.6	27.5	21.7	21.6	22.0	28.8
							✓	42.4	41.8	44.4	56.6	54.6	48.0	49.3	54.5	49.0
						✓	✓	52.1	56.7	59.5	63.0	65.7	60.5	59.5	57.3	59.3
				Ortho.	✓			72.6	79.6	72.1	79.1	66.9	76.6	73.8	80.5	75.2
						✓		72.0	73.6	76.7	77.3	83.1	75.4	72.4	75.5	75.7
							✓	**86.3**	76.6	71.3	79.6	**88.8**	**81.7**	75.4	**85.7**	80.7
						✓	✓	85.5	**81.6**	**76.8**	**82.4**	88.2	81.2	68.6	81.8	**80.8**

Table 3.3: Comparison with prior studies on the collected dataset under the sole cross-view condition (%)

Networks	Modalities	Projections	Viewpoints			Mean
			Sensor-view	Side-view	Back-view	
Benedek *et al.* [5]	GEI	Spher.	✓			62.6
				✓		71.7
		Ortho.	✓			71.9
				✓		76.5
Shiraga *et al.* [61]	GEI	Spher.	✓			70.9
				✓		71.1
		Ortho.	✓			71.9
				✓		76.5
Yamada *et al.* (Network 1) [79]	Depth Seq.	Spher.	✓			53.4
				✓		49.8
		Ortho.	✓			64.1
				✓		62.1
Yamada *et al.* (Network 2) [79]	Depth Seq.	Spher.	✓			44.1
				✓		54.0
		Ortho.	✓			54.9
				✓		61.2
Ahn *et al.* [1]	Depth Seq.	Spher.		✓	✓	93.4
		Ortho.		✓	✓	**94.8**
Proposed	Depth Seq.	Spher.	✓			89.6
				✓		54.6
					✓	83.4
				✓	✓	89.9
		Ortho.	✓			89.7
				✓		86.1
					✓	91.0
				✓	✓	93.8

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

Modalities		Temporal Encoding (TE)		Mean
Silhouette Seq.	Depth Seq.	1D-LSTM	ConvLSTM [59]	
✓		hidden size = 256		49.2
✓		hidden size = 512		58.4
✓		hidden size = 1024		57.6
✓			kernel size = 3×3	69.7
✓			kernel size = 5×5	67.1
✓			kernel size = 7×7	66.2
	✓	hidden size = 256		51.8
	✓	hidden size = 512		65.2
	✓	hidden size = 1024		65.9
	✓		kernel size = 3×3	**72.1**
	✓		kernel size = 5×5	70.4
	✓		kernel size = 7×7	68.5

Modality and TE

I first investigated the effectiveness of the proposed network by changing the input modalities, temporal aggregating manners, and hyper-parameters, as presented in Table 3.4. In this case, I conducted the experiment using only a single original view without the VE module because applying 1D-LSTMs in the TE module that includes the ATFF block, which addresses 2D spatial features, is difficult. Silhouettes are commonly used as the input format for gait recognition using RGB cameras. In contrast, depth images are obtained from 3D depth sensors, including LiDAR sensors, which contain richer contextual information than silhouettes and RGB images. In the TE part, I enhanced the recognition performance by replacing the previously used LSTMs in [79] with ConvLSTMs [59] to effectively extract spatio–temporal features. The accuracies shown in Table 3.4 represent the averages across all the cross-view and cross-distance combinations. The use of depth image sequences and ConvLSTMs yields superior results compared with others, either individually or in combination. This improvement could be attributed to two main reasons: depth images capture gait features better because of their textural information, and ConvLSTMs consider temporal features with an additional spatial property and learn them simultaneously, as opposed to 1D-LSTMs.

Impact of RE

In Table 3.5, the experiment was conducted using only a single spatial encoding unit of the VE module with the original view, to evaluate the effectiveness of the RE module with the ATFF block. This setting is the same as that in the aforementioned ablation study. In Table 3.5, the comparison metrics $\mathbf{I}_{\text{high-res}}$, $\hat{\mathbf{I}}_{\text{low-res}}$, $\hat{\mathbf{f}}_{\text{high-res}}$, $\hat{\mathbf{f}}_{\text{low-res}}$, and $\mathbf{f}_1$ are presented in Fig. 3.4 (b) and Fig. 3.6. The average accuracies obtained using a single resolution were lower than the results in the last four rows of Table 3.5, using only a single experiment by changing the components of the VE module

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

and the spatial encoding unit of the VE module with the original view. Gait images with low resolution are insufficient for independently extracting spatial features. Among the fusion methods, the proposed ATFF block with squeezing the dimension T achieved the highest performance and outperformed both the simple combination of two spatial features and the direct application of the original architecture from [22]. This performance could be attributed to the ATFF block re-calibrating the scores of both input features in a correlated manner, which resulted in considerable robustness and adaptability to changing conditions. Compared with the original structure of [22] that considers channel-wise interrelationships, in the proposed attention strategy, the two spatial features are fused by mutually comparing their scores across channels. Furthermore, the use of the ATFF block with T pooling compresses global temporal features related to gaits, which results in improved discrimination power compared with not using temporal pooling. In the last two rows of Table 3.5, a slight difference in recognition accuracy was observed when the two targets $\hat{\mathbf{f}}_{\text{low-res}}$ and $\mathbf{f}_{\text{high-res}}$ were switched in the ATFF block. This result indicates that The two scores $\mathbf{s}_1$ and $\mathbf{s}_2$ converge to be the same during the training.

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**

Table 3.6 presents the results of the ablation experiment to investigate the effect of the VE module, which utilizes two invariant gait shapes with the ATFF block to fuse their features. This experiment was conducted by changing the components of the VE module based on the complete

proposed network. In the first three lines of Table 3.6, the performance adopting the GDT process outperforms that of the original view. This phenomenon suggests that aligning the walking directions allows for better extraction of coherent gait patterns. In the overall average accuracies in Table 3.6, the approaches that consider two invariant viewpoints exhibit superior performance compared with those of single views. This method demonstrates better performance than either pooling approaches, which were used in the previous version [1], or the simple concatenation approach. These results indicate that the ATFF method renders the proposed network more robust to self-occlusions caused by changes in the emission direction of the lasers. This effectiveness was achieved by considering the higher influence between the two viewpoints in an end-to-end manner.

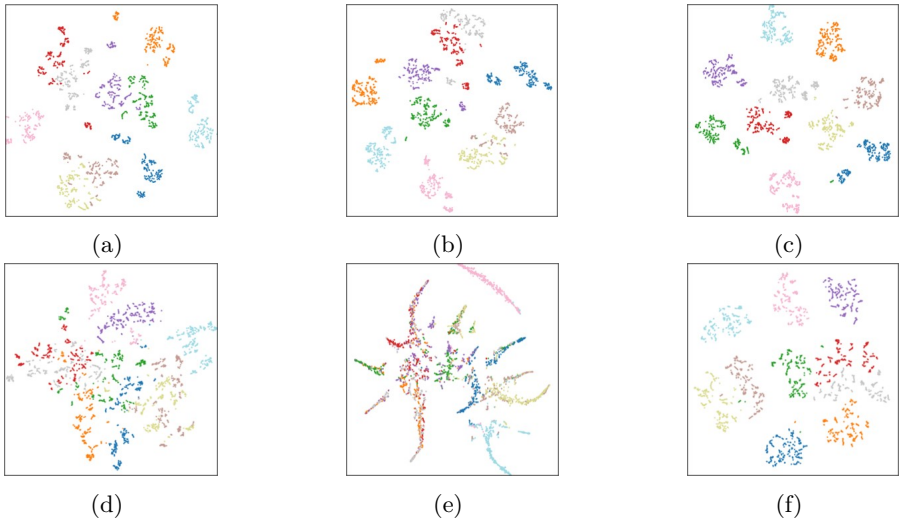
3.5.5 Practicality

Galleries collected from practical scenarios have limitations in terms of viewing angles or quantity of data, compared with typical cross-view challenges. The proposed recognition model could be more effective in these limited conditions because it flexibly uses the gait shapes of the two viewpoints. In this section, I investigated the practicality of the proposed model, comparing with the prior methods [6, 61, 79]. The experiment was conducted by restricting the viewing angle of the galleries. Specifically, I saved only a single angle as a database for each of the following: 270° (side view), 0° (back view), and 315° (oblique view). Compared with typical cross-view experiments, this scenario is more challenging because of limitations not only in viewing angles and gait sequences but also in the point cloud densities of the galleries. I evaluated the recognition models from three perspectives, namely projection, viewpoint, and network, for which the combinations are the same. Each accuracy value in Table 3.7 is the average of eight probe views and two cross-distances per viewing angle of the gallery.

In Table 3.7, the proposed projection approach (Ortho.) outperformed the previous way (Spher.) for all networks in terms of the original view. A possible reason for this is that the sorted walking directions representing consistent dynamics can extract the gait features. When the viewing angles of the galleries and viewpoints transformed using the GDT processing were identical, this orthographic projection improved the recognition performance considerably, except for Network 1 in [79]. When the walking angles of the galleries reached the target angles of the GDT, visual differences such as partial occlusions or depth values in gait images were observed, which resulted in distinct gait features. The accuracies of the prior spherical projection deteriorated when the GDT was applied. Based on these results, the point cloud projection approach achieves superior compatibility with the transformation of gait angles because of the undistorted geometric features of the gait. Finally, the proposed model, which utilizes two fixed gait shapes selectively with the proposed attention manner, achieved the best performance for all combinations, from all viewing angles of the galleries.

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.* [61]	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) [79]	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) [79]	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
Proposed	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**



(a) (b) (c)

(d) (e) (f)

Figure 3.9: t-SNE visualization of the gait features from 10 subjects, each with 8 views, 2 distances, and 42 sequences. The top row shows the proposed model, in which the RE and VE modules are applied in order from left to right. In this case, (a), (b), and (c) correspond to the top line of Table 3.5 and the top and bottom lines of Table 3.6, respectively. The bottom row shows the prior methods, with [61], [79], and [1] are listed in order from left to right. In this case, all these networks are applied to the proposed point cloud projection and GDT processing.

3.5.6 Feature Visualization through t-SNE

In this section, I describe a qualitative evaluation by applying t-SNE [72] to visualize gait features through a 2D manifold space. Using the learned recognition models, I extracted features from gait sequences for all the eight viewing angles and two distances with ten subjects who were not used in the training. The top row in Fig. 3.9 presents a visualization of gait features extracted from the proposed model, which were gradually applied with RE and VE modules from left to right. Adding more proposed modules results in narrower intra-class distances and wider inter-class margins, as displayed in Fig. 3.9. I visualized the features of the prior methods in the bottom row of Fig. 3.9, representing the results of [61], [79], and [1] from left to right. Here, these recognition networks were applied to the proposed point cloud projection (Ortho.) and the GDT processing because these approaches achieved the best accuracy for each network in this section. The distances of points between intra- and inter-class reveal that the proposed complete model exhibits superior discrimination power, compared with the three prior methods.

3.6 Conclusion

I proposed a depth-based gait recognition model using a LiDAR sensor, which is a modified version of the model presented in the previous chapter. The effectiveness of the model was explored in-depth from three perspectives: point cloud projections, gait direction transformation, and the gait recognition network. In this model, gait shapes from multiple invariant viewpoints are generated from point cloud sequences, and gait features are extracted using a proposed attention method that effectively fuses two similar features. Experiments on the collected dataset demonstrated that the proposed approach achieved superior recognition performance in both cross-view and cross-distance challenges compared to prior methods. Moreover, based on extensive experimental results, the proposed model exhibited significant potential for practical applications, even in scenarios with limited viewing angles.

Part II: The Development of Gait Upsampling Models

Chapter 4

Upsampling Modeling for LiDAR Gait Sequence Data

4.1 Introduction

In the previous chapter, I introduced a gait-based identification model utilizing LiDAR projection to address sparse LiDAR gait data. Through comprehensive experiments, I validated its potential applicability in long-range measurement environments. However, this approach requires training the network with sparse gait inputs, which are influenced by variations in both the range and emission specification of LiDAR sensors. Collecting a comprehensive training dataset that accounts for all possible cases of gait sparsity is impractical. Furthermore, sparse gait data inherently lacks fine-grained appearance information, making it challenging to achieve accurate person identification using only a recognition network.

Recent studies on 3D LiDAR-based gait recognition have emerged, enabled by the large-scale LiDAR gait dataset, *SUSTeck1K* [58]. These studies demonstrate that 3D LiDAR sensors outperform traditional RGB cameras due to their independence from lighting conditions and their ability to capture precise 3D geometry. However, these findings are limited to short-range scenarios (5-15 m) and rely on high-specification, high-cost sensors, such as the 128-beam LiDAR scanner (Velodyne VLS-128), [18, 58]. When deploying LiDAR sensors in person identification systems, the sparsity of pedestrian data is highly influenced by measurement distance and hardware specifications, particularly with low-resolution sensors such as 32-beam scanners discussed in Chapter 2 and 3. This sensitivity often leads to significant degradation in identification performance due to sparse or incomplete representation of the human body, limiting the ability to capture the fine-grained structure of the original pedestrian shapes. Addressing these difficulties requires reconstructing the underlying or the complete gait shapes from sparse LiDAR data.

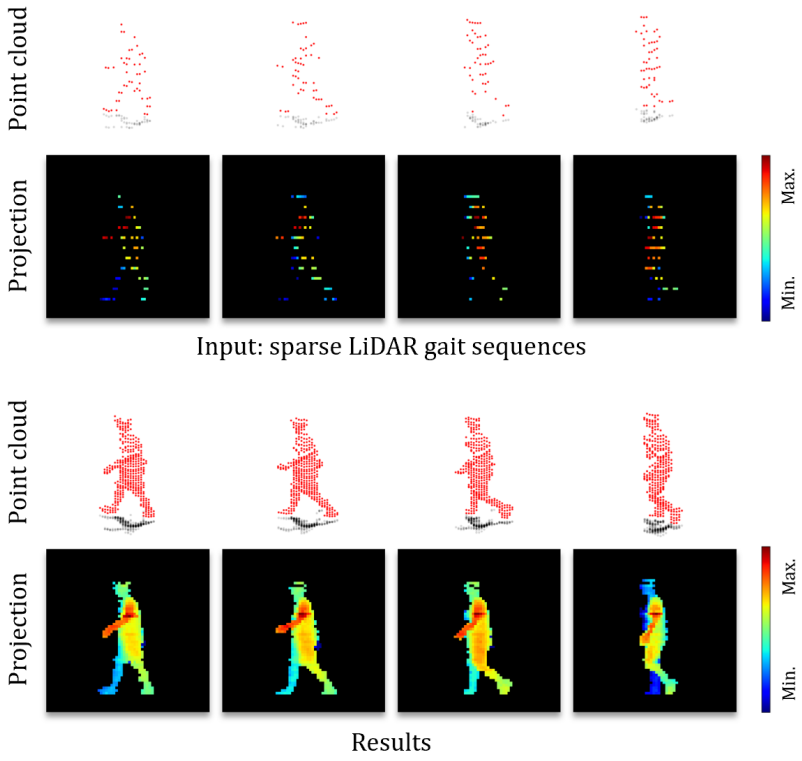


Figure 4.1: Upsampled results using the proposed model. I present sparse LiDAR gait sequence data as inputs (top two rows) alongside the corresponding outputs (bottom two rows), represented in both 3D point cloud sequences (rows 1 and 3) and 2D depth videos (rows 2 and 4).

To address this challenge, in this chapter, I introduce a gait sequence upsampling model for sparse LiDAR pedestrian data, as shown in Fig. 4.1, aiming to enhance the generalization capability of existing identification models. The proposed method utilizes diffusion probabilistic models (DPMs) [20], which have shown high fidelity in generative tasks, including image completion. Specifically, I 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, I employed a video-based noise prediction model [27] during the denoising process. To the best of my knowledge, this is the first study to address LiDAR data upsampling for gait recognition. In the experiments, I demonstrate the effectiveness of the proposed model on two datasets: the *SUSTeck1K* [58], a large-scale gait dataset, and the collected dataset used in Chapter 3. Both datasets were utilized to evaluate generative quality and improvements in identification performance.

The contributions of this study can be summarized as follows:

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

4.2 Related Work

4.2.1 Gait-based Identification Models

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 [9, 12]. In contrast, the latter parameterizes visual data into human structures, such as shape-aware and non-shape-aware poses [69, 87], 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.* [5, 6, 16] proposed a projection-based model using gait energy images (GEIs) [17] to re-identify individuals in short-term scenarios. However, this method struggles to satisfactorily extract dynamic features from gait frames. The previous study [79] 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. Building on these challenges, I explored view- and resolution-robust recognition frameworks by rearranging pedestrian point clouds based on a gait direction vector [1, 2]. Although this method enhances identification performance under complex confounding conditions, it still lacks the geometric features necessary for distance-independent analysis. Shen *et al.* [58] 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 [12] 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 [79]. Han *et al.* leverage two modalities, including range view images and raw point clouds from LiDAR data, to utilize essential geometric information.

4.2.2 Diffusion Probabilistic Models for LiDAR Data Generation

Diffusion-based generative models [20, 62, 65] have gained significant attention across a wide range of applications, including text-to-image generation [53], speech synthesis [29], translation [36], and compression [68]. Compared with generative adversarial networks (GANs) [14], 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) [20], a type of diffusion-based model, have demonstrated notably high fidelity for various completion tasks [52, 54]. 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.* [88] utilized NCSNv2 [64], a score-based generative model, to train both the range and reflectance modalities using image representations. Nakashima *et al.* [48] adopted unconditional DDPMs, incorporating inpainting and timestep-agnostic techniques [27,43], to enhance both the fidelity and efficiency of LiDAR data synthesis for sim2real applications. Sander *et al.* [19] introduced LiDAR upsampling models based on conditional DDPMs [54], achieving faster sampling while maintaining high fidelity compared with prior works [48,88]. In contrast to [19], which utilizes spherical projection, I employ conditional DPMs [54] for LiDAR gait data completion, regardless of the sensor's distance, using an orthogonal projection strategy. Furthermore, I design a video-to-video translation model to ensure the time-sequential consistency of the gait data.

4.3 Method

In this section, I 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.

4.3.1 Problem Statement

The 3D point cloud data captured from single LiDAR sensors can generally be transformed into range images. Most existing studies [19, 48, 88] 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 directly maps LiDAR point clouds onto depth images within 3D Euclidean space. Compared to spherical projection, the orthographic projection method [1, 2, 6] 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 [12, 58]. 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:

$$\mathbf{y} = \boldsymbol{H} \odot \mathbf{x}_0 + \epsilon, \tag{4.1}$$

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, I assume that the noise ϵ is set to zero. My 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.

4.3.2 LiDAR Data Representation

Based on the problem statement, I 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 $\mathcal{P}_i^j$, 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 is defined 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 only the sensor emission directions on the xy-plane are considered. 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 $\mathcal{P}_i^j$, the rotated point cloud sequence $\hat{\mathcal{P}}_i^j \in \mathbb{R}^{F \times N \times C}$ with a directional angle $\theta_{\text{sensor}}{}_{i,f}^{j}$ is obtained as follows:

$$\hat{\mathbf{p}}_{i,f,n}^j = (\mathbf{p}_{i,f,n}^j - \mathbf{c}_{i,f}^j) \cdot \mathbf{R}_z(\theta_{\text{sensor}}{}_{i,f}^{j} + \pi), \tag{4.2}$$

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

As in the chapter 3, the point cloud sequence $\mathcal{P}_i^j$ is transformed 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:

$$h = \left\lfloor \frac{1}{r_z} \cdot (\hat{p}_{i,f,n,z}^j - \min_{n \in \{1,...,N\}} (\hat{p}_{i,f=0,n,z}^j) + l_{z-\text{const}}) \right\rfloor, \tag{4.3}$$

$$w = \left\lfloor \frac{1}{r_y} \cdot (\hat{p}_{i,f,n,y}^j + \frac{l_y}{2}) \right\rfloor, \tag{4.4}$$

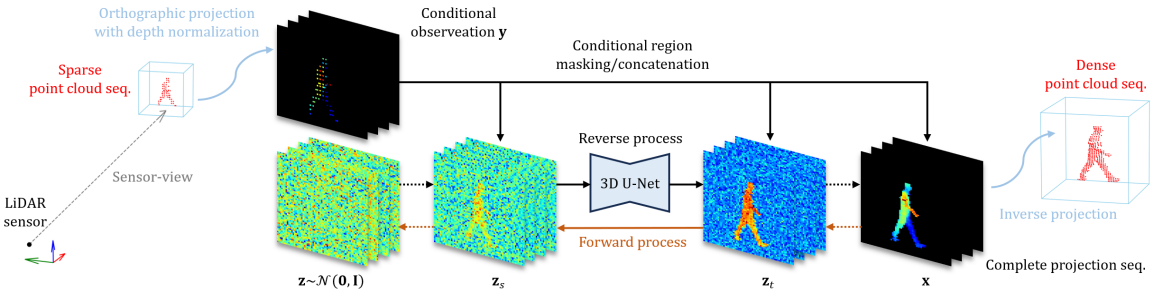


Figure 4.2: Overview of the upsampling pipeline. The diffusion processes operate within the orthographic projection domain with normalized depth values. The sampled depth projection sequences are then translated into 3D point cloud data.

, 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.

4.3.3 Preliminaries for DDPMs

In this study, inspired by [54], I build a diffusion-based inpainting model, as shown in Fig. 4.2, conditioned on observation $\mathbf{y}$ (for simplicity, j and i are omitted). Additionally, I employed the DDPM framework, which formulates transitions between data and latent spaces with continuous time $t \in [0, 1]$ [27]. Compared with discrete-time diffusion models [20], 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:

$$q(\mathbf{z}_t|\mathbf{x}_0) = \mathcal{N}(\alpha_t \mathbf{x}_0, \sigma_t^2 \mathbf{I}), \tag{4.5}$$

where α_t and σ_t^2 are strictly positive scalar-valued functions of t in the noise schedule. In this study, I employed α-cosine schedule [49], 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 [65]. 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:

$$q(\mathbf{z}_t|\mathbf{z}_s) = \mathcal{N}(\alpha_{t|s}\mathbf{z}_s, \sigma^2_{t|s}\mathbf{I}), \tag{4.6}$$

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

$$p(\mathbf{z}_s|\mathbf{z}_t) = \mathcal{N}(\mu_t(\mathbf{x}_0, \mathbf{z}_t), \Sigma^2_t\mathbf{I}), \tag{4.7}$$

where $\mu(\mathbf{x}_0, \mathbf{z}_t) = \frac{\alpha_{t|s}\sigma^2_s}{\sigma^2_t}\mathbf{z}_t + \frac{\alpha_s\sigma^2_{t|s}}{\sigma^2_t}\mathbf{x}_0$ and $\Sigma^2_t = \frac{\sigma^2_{t|s}\sigma^2_s}{\sigma^2_t}$.

4.3.4 Noise Prediction Model

In this study, I used a modified 3D U-Net architecture [27] 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 [20, 54], 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.

4.3.5 Loss Function

I define the objective function for the proposed 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:

$$\mathbf{z}_t \leftarrow \mathbf{m} \odot \mathbf{y} + (\mathbf{1} - \mathbf{m}) \odot \mathbf{z}_t, \tag{4.8}$$

$$m_{(f,1,h,w)} = \begin{cases} 1, & \text{if} \quad y_{(f,1,h,w)} > 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.9}$$

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:

$$\mathcal{L}_{T\to\infty} = \mathbb{E}_{\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I}),t\sim\mathcal{U}(0,1)}[||\hat{\epsilon}(\text{concat}(\mathbf{y}, \mathbf{z}_t); \lambda_t) - \epsilon||^2_2]. \tag{4.10}$$

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, I mask the loss to compute only the unknown regions in the depth videos for more efficient training, as in [19].

4.4 Datasets

The performance of the proposed model was evaluated using two datasets. The first is the *SUSTeck1K* dataset [58], 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. Compared to the datasets used in Chapters 2 and 3, which consist only of the *Normal* walking condition, this dataset involves 12 gait attributes, such as *Bag* and *Umbrella*. In addition, this dataset includes data from 1,050 identities and eight viewpoints, making it suitable for training both identification and generative models, as well as for evaluating general-purpose performance. In the experiments, I used the training subset of the *SUSTeck1K* as a clean/complete dataset to learn the distribution of gait shapes for upsampling models.

The second dataset [2] was collected in previous Chapter 4 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 4.1. 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).

Table 4.1: Comparison between two datasets

Datasets	Sensors	Beams	V/H Resolutions	Subjects	Views	Distances
SUSTeck1K [58]	VLS-128	128	0.11°/0.1°	1,050	12	7.5 m
Ahn *et al.* [2]	VLP-32C	32	1.33°/0.1°	30	8	10, 20 m

4.5 Experiments

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

4.5.1 Implementation Details

Following the original protocol, I used a subset of *SUSTeck1K*, consisting of 250 subjects, for training the upsampling model. I trained the proposed model for 200,000 iterations with a learning rate of 0.0003 and a 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. The code for all experiments

was implemented using Python in PyTorch 2.0, and executed on a dual NVIDIA GeForce RTX 3090 GPU (approximately 45 hours to train). I 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. 4.3. 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, I used three different ratios for each noise mask type and paired them with the original gait data during the training.

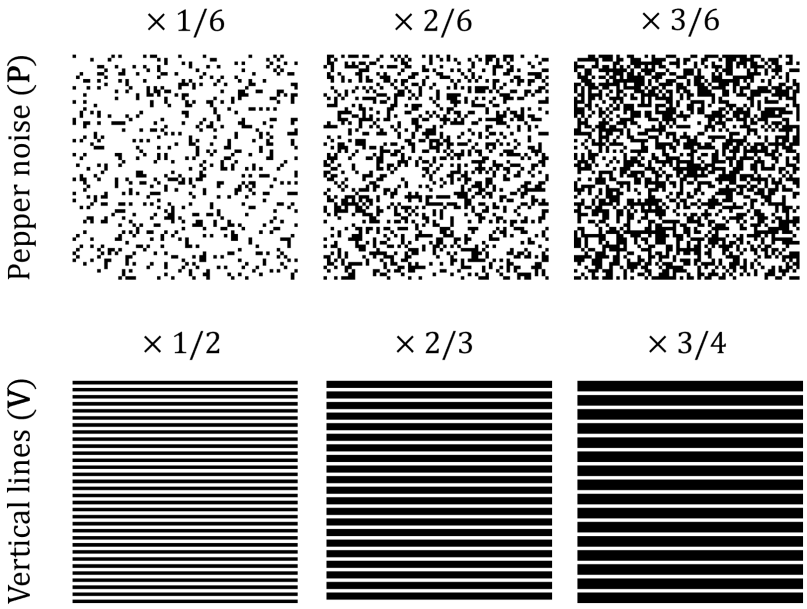


Figure 4.3: Noise masks used for training and testing the proposed model. All mask sizes are 64×64, and the black regions in each binary noise mask indicate the points removed from the clean gait data.

For the testing phase, I used the remaining test set of *SUSTeck1K*, which consists of 800 subjects, randomly selecting 10 frames for each with three different combinations of noise masks. In the applicability experiment, I used the dataset collected in Chapter 3 [2]. As a baseline, I compared the proposed diffusion model with the vanilla Palette [54] using the proposed projection on two datasets. In addition, I compared the well-established linear interpolation methods: Nearest-neighbor, Bilinear, and Bicubic. In the generative quality evaluation, I adopted the following two widely-used standard metrics: the Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). I 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, I used LidarGait [58] 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 the proposed upsampling model and LidarGait, trained on the training set of *SUSTeck1K*, were applied in all experiments.

4.5.2 Generative Evaluation

The quantitative generative evaluation results on the *SUSTeck1K* dataset are listed in Table 4.2, and the examples sampled by the proposed model for the gait attribute *Normal* are shown in Fig. 4.4. In Table 4.2, diffusion-based methods significantly outperformed the interpolation approach across all three metrics. Comparing the proposed model to Palette [54], I observed that the video-based model [27] is more effective than the image-based approach. Notably, as the noise masks became more severe, the performance gap between the proposed model and Palette increased. For other gait attributes in Fig. 4.5, it can be observed that the proposed method realizes high fidelity in both 2D projected gait shapes and the structure of 3D point clouds.

Table 4.2: 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 [54]	Depth Image	26.14	0.940	0.009	24.17	0.908	0.013	23.15	0.888	0.017
Diffusion	Proposed 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	Proposed	Depth Video	**27.27**	**0.954**	**0.007**	**25.59**	**0.932**	**0.010**	**24.89**	**0.922**	**0.011**

4.5.3 Gait Recognition Task

The identification results conducted on *SUSTeck1K* using LidarGait [58] are listed in Table 4.3. In Table 4.3, it can be observed that the interpolation approach achieves little to no improvement in recognition performance on sparse gait data. Similar to Table 4.2, it can be observed that the proposed 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. 4.6.

Fig. 4.7 shows the identification evaluation scores as functions of the number of function evaluations (NFE), which 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.

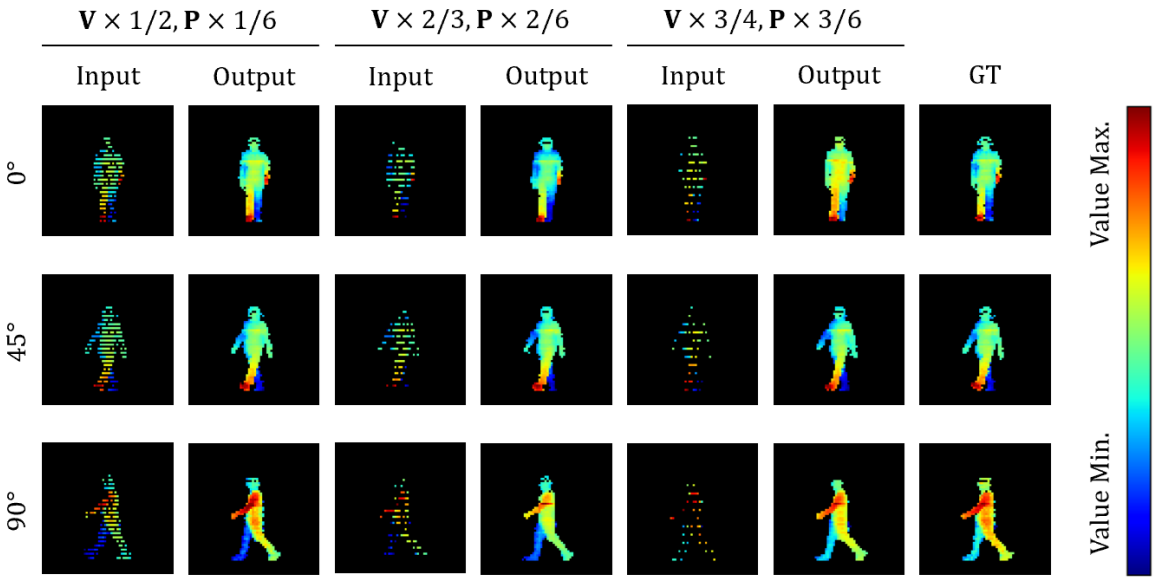


Figure 4.4: Upsampled results using the proposed model on the *SUSTeck1K* dataset for the *Normal* attribute with three noise mask combinations. The pixels represent depth values calculated from behind in the sensor's emission direction with depth normalization, where red indicates greater depth, as shown in the color bar on the right. In other words, the redder the color, the closer it is to the sensor. This representation is consistent across all figures in the projection domain represented in Chapter 4.

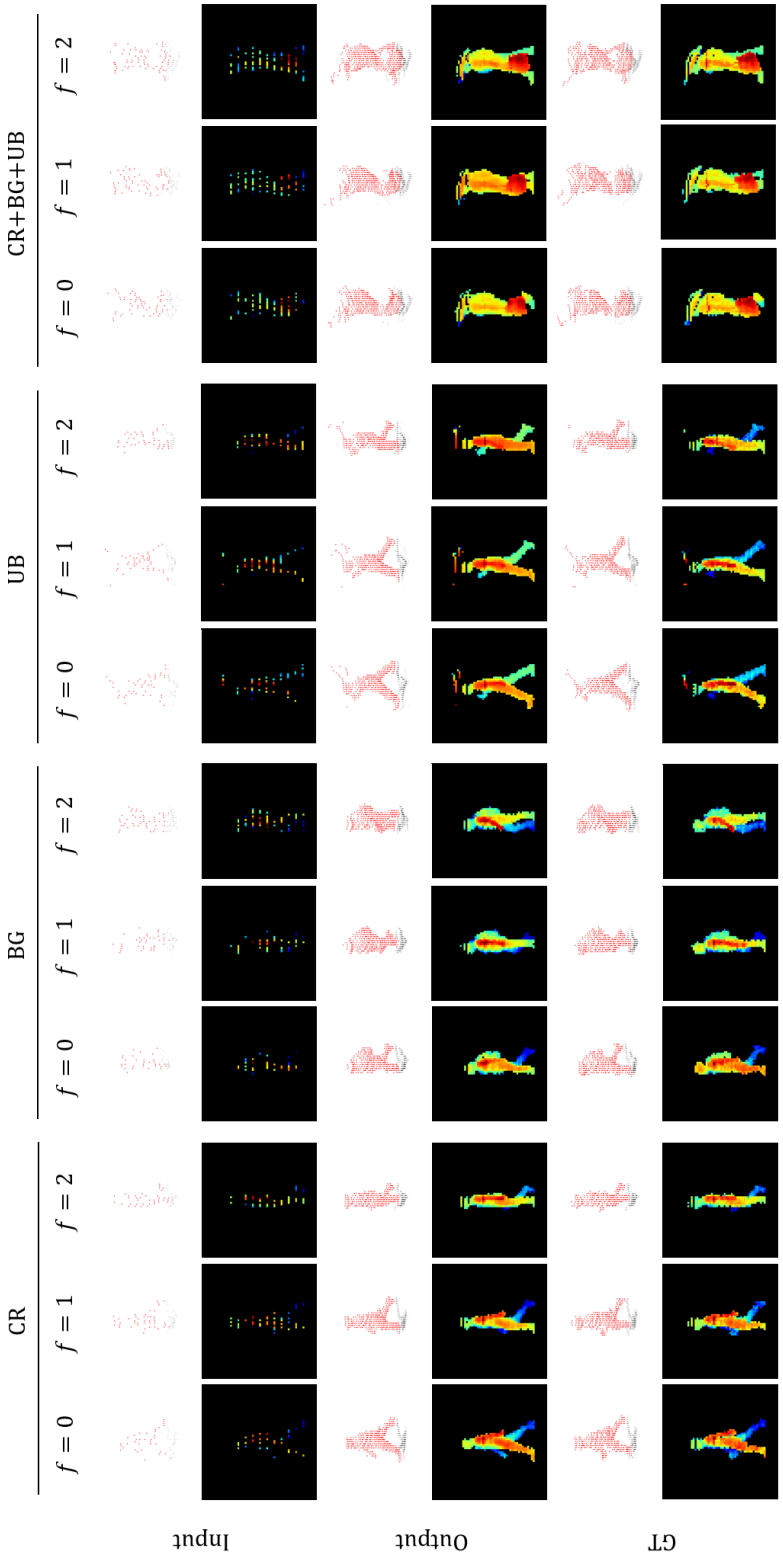


Figure 4.5: Upsampled results using the proposed model from noise masks with $\mathbf{V}\times 3/4$ and $\mathbf{P}\times 3/6$. I showcase the samples for three gait variances: Carrying (*CR*), Bag (*BG*), and Umbrella (*UB*).

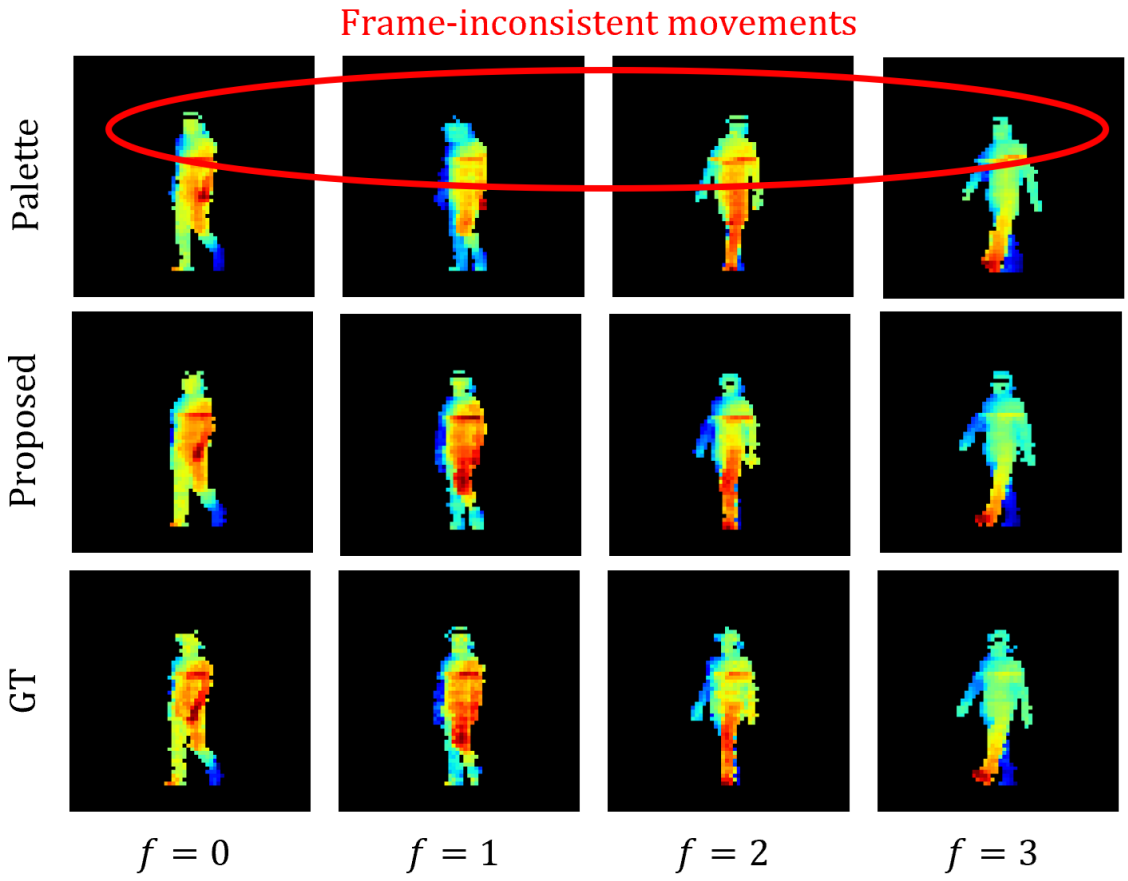


Figure 4.6: Comparison between the proposed model and Palette [54]. The results are sampled from noise masks with $\mathbf{V}\times 3/4$ and $\mathbf{P}\times 3/6$ (top two rows).

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 [54]	Depth Image	23.62	48.69	61.07	9.93	26.61	37.31	7.16	13.79	21.82
Diffusion	Proposed 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	Proposed	Depth Video	**32.49**	**59.77**	**71.28**	**18.97**	**42.09**	**54.52**	**11.85**	**30.68**	**42.26**

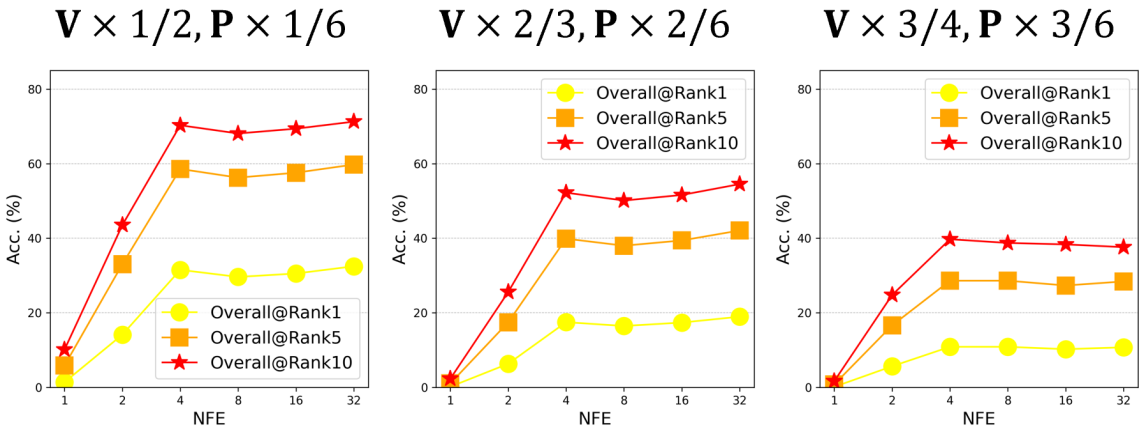


Figure 4.7: Comparison of the number of function evaluations (NFE) for the proposed model by sweeping T across $\{1, 2, 4, 8, 16, 32\}$.

4.5.4 Application

The identification results conducted on the collected dataset [2] using LidarGait [58] to evaluate the applicability of our model are shown in Table 4.4. In addition, the gait shapes according to the two different point cloud projections as illustrated in Fig. 4.8. In Table 4.4, I observed that our upsampling method 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.

4.6 Conclusion

I introduced an upsampling method for LiDAR-based gait sequence data to address the distance-independent inpainting problem. A conditional diffusion model was employed to tackle the inpainting problem in gait videos. Through comprehensive experiments, the proposed upsampling model demonstrated significant performance improvements over existing methods in terms of both generation quality and gait recognition. Notably, the proposed model proved effective even for pedestrians

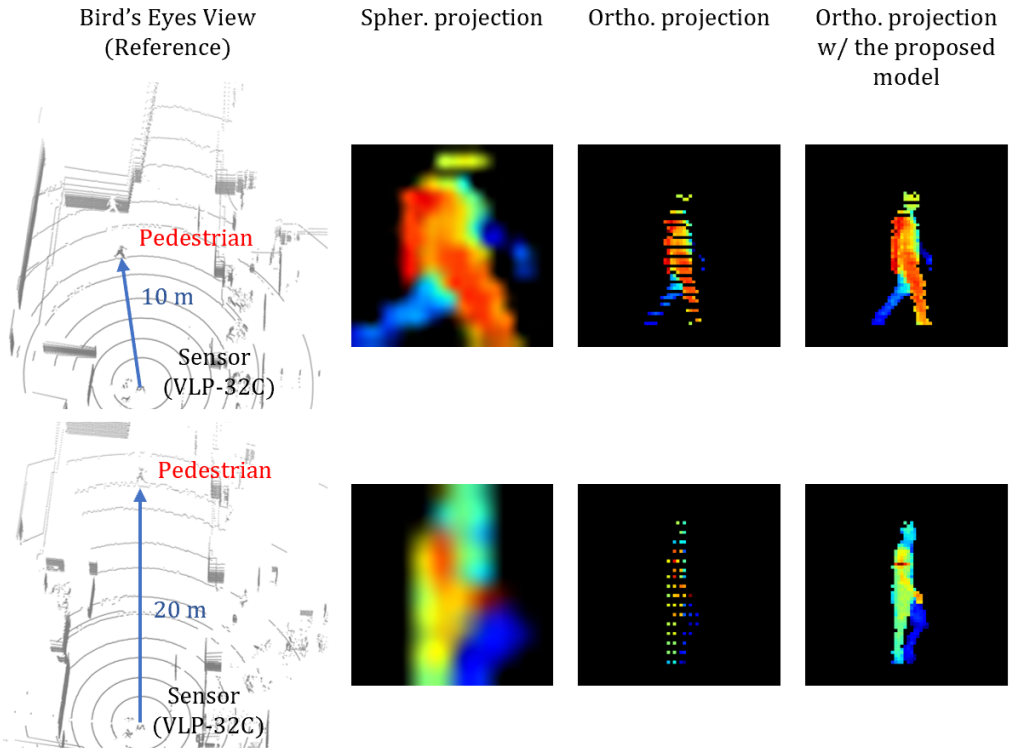


Figure 4.8: Projection comparison on the real-world dataset [2] at two capture distances: spherical projection (Spher.) and the proposed orthographic projection (Ortho.).

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 [54]		✓	Ortho.	19.57	56.25
	✓	✓	Ortho.	25.45	63.54
Proposed		✓	Ortho.	21.28	60.94
	✓	✓	Ortho.	**25.97**	**66.82**

with varying sensor resolutions or measurement distances in real-world scenarios.

Chapter 5

Conclusion

In this dissertation, I have presented deep learning-based approaches in two aspects to enhance the identification performance of 3D LiDAR-based gait recognition.

In Part I, I introduce a gait recognition model designed to be robust against variations in walking direction and measurement distance. Compared to previous studies, this model employs orthogonal projection to reduce errors caused by linear interpolation preprocessing and to accurately represent pedestrian size in gait videos. As a gait representation, gait shape sequences from two invariant viewpoints, combined with depth normalization and speed sequences, are employed to enhance discriminative capability by leveraging the unique characteristics of 3D LiDAR sensors. For the network technique, a multi-scale spatial encoding strategy is implemented to mitigate the adverse impacts caused by changes in the sparsity of gait shapes. In the modified version of the above model, an attention-based feature fusion strategy led to performance enhancement. In the experiments, the effectiveness of the proposed model was demonstrated using datasets collected from two different types of LiDAR sensors.

Despite the advantages of the proposed identification models presented in Part I, several research questions remain unanswered. First, self-occlusion inherently occurs when projecting 3D pedestrian point clouds in a specific direction. How can partial gait shapes be effectively learned to match the gait features of complete gait data? As a potential approach to address this, it is considered feasible to leverege generative models or foundation models to utilize the underlying distribution of complete gait data for recovery and identification. Second, when pedestrian data becomes sparse, speed information provides meaningful assistance; however, in the proposed model, there is a tendency for identification performance to slightly decrease at close distances. How can pedestrian speed information be effectively learned by the network as a gait feature? The walking speed information used in Chapter 2 is considered to have oversimplified pedestrian movement on the xy-plane, which is likely the reason for the performance degradation at short distances. Currently, research is being conducted on reconstructing and modeling the human body using 3D LiDAR [33]. If the human

skeleton can be accurately inferred from sparse LiDAR data, it is believed that the movement or trajectory of the human pose over time can be effectively utilized as a gait feature. Collectively, these questions and expectations shed light on future research directions for gait identifier modeling.

In Part II, I introduce an upsampling model designed to perform LiDAR sparse-to-dense tasks for gait sequence data. Especially, it is intended to enhance the generalizability of identification models for long-distance measurements. In this model, gait data is processed using a distance-independent inpainting by orthogonally projecting gait data as viewed from the LiDAR sensor. Additionally, video-based conditional diffusion models were developed to ensure the consistency of pedestrian movements over time. In the learning phase, a noise mask strategy is employed to account for changes in LiDAR data sparsity based on measurement distance. In the experiments, the proposed upsampling model outperformed traditional interpolation methods and vanilla restoration methods on both simulated noise datasets and real-world datasets with different types of sensor hardware. Notably, experiments demonstrate that even for identification models trained exclusively on complete gait data, the proposed upsampling model significantly enhances performance for gait data captured by low-resolution LiDAR sensors or at long-range measurements.

The above practice demonstrates the effectiveness of the proposed model in Part II; however, several issues remain. First, the diffusion model employed in this approach involves an iterative generation process, resulting in slow processing speeds and high computational costs, which are impractical for real-time person identification applications. How can the inference speed for upsampling gait data be improved? Recently, Flow Matching (FM) [40] has emerged as a generative technique that matches the probability flow field between a source and a target distribution by solving an ordinary differential equation (ODE). Unlike DPMs, it requires fewer steps for generation due to its deterministic and noiseless ODE framework, which makes the sampling process more efficient. Applying the FM technique to upsampling models for LiDAR gait data is expected to reduce inference time. Second, there is a growing need to address gait data affected by more complex noise patterns, such as obstacle occlusion, shifting the focus toward gait data restoration rather than simple upsampling. How can we address multi-task gait data restoration that includes occlusion? In Chapter 4, I proposed a conditional diffusion model that incorporate task-specific approaches [52, 54]. In contrast, task-agnostic approaches learn the underlying data distribution and reconstruct data through posterior sampling [43, 63]. While it can be easily plugged into various tasks without additional training, their performance tends to be worse than task-specific approaches. If a restoration model capable of effectively learning video or sequential point cloud data is developed, it is expected to address not only upsampling but also challenges such as obstacle occlusion or frame-drops. These suggestions will serve as a foundation for guiding future research directions.

The contributions of this study enable new capabilities, some of which are listed below.

Domain extension for gait recognition

In the field of gait recognition, RGB cameras have traditionally been the primary devices for collecting pedestrian data. On the other hand, LiDAR has emerged as a new technology that complements the shortcomings of camera-based gait recognition, such as sensitivity to lighting conditions or viewing angles. However, the sparse gait data captured by these LiDAR sensors is known to be the biggest challenge for person identification, which has limited its application in real-world environments. By dealing with this challenge with two different perspectives—identification and upsampling models—we not only take a step closer to practical applications but also inspire future research in LiDAR-based gait recognition. Specifically, since this study leverages a LiDAR projection strategy, it is expected to demonstrate strong compatibility with camera-based gait recognition, which addresses 2D videos of walking silhouettes (cross-domain).

Impacts of computer vision

In computer vision, LiDAR sensors have primarily been used to address tasks such as scene understanding in autonomous driving. However, few advances have been made in tasks related to single objects, such as humans, because of the low density of LiDAR data. I believe that this study, which addresses sparse human data, opens new horizons for similar tasks such as object recognition using LiDAR sensors. On the other hand, the study proposed in Chapter 4 is expected to inspire restoration tasks, especially those dealing with sequential data (e.g. videos), which have made very little progress. Furthermore, it is not limited to upsampling gait data and has potential applicability to other 3D human restoration tasks involving diverse poses.

Societal impacts

In terms of impacts of society, this study is expected to advance person identification capabilities using LiDAR technology. Particularly, it is expected to enable the additional functionalities in applications where LiDAR is essential, such as mobile robots or autonomous vehicles. For example, LiDAR-based gait analysis could perform various functions, such as user identification and disease diagnosis. Furthermore, more accurate gait analysis tasks could be achieved by moving beyond traditional models that handle 2D images and instead utilizing 3D human modeling.

Research on 3D LiDAR-based gait recognition has only recently begun to make progress but still lags behind camera-based research. Building on this study, I believe it will contribute not only to advancements in person identification using LiDAR sensors but also to the realization of practical applications in the future.

Chapter 6

Ethical Approval

This study was approved by the Ethics Committee of the Graduate School of Information Science and Electrical at Kyushu University.

Bibliography

[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] Jeongho Ahn, Kazuto Nakashima, Koki Yoshino, Yumi Iwashita, and Ryo Kurazume. Gait sequence upsampling using diffusion models for single lidar sensors. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, 2025.

[4] Michal Balazia and Petr Sojka. You are how you walk: Uncooperative mocap gait identification for video surveillance with incomplete and noisy data. In *Proceedings of the IEEE International Joint Conference on Biometrics (IJCB)*, pages 208–215, 2017.

[5] Csaba Benedek. 3d people surveillance on range data sequences of a rotating lidar. *Pattern Recognition Letters*, 50:149–158, 2014.

[6] 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.

[7] Imed Bouchrika, Michael Goffredo, John N. Carter, and Mark S. Nixon. On using gait in forensic biometrics. *Journal of Forensic Sciences*, 56, 2011.

[8] Rich Caruana, Steve Lawrence, and C. Giles. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 13, 2000.

[9] 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.

[10] Pratik Chattopadhyay, Shamik Sural, and Jayanta Mukherjee. Frontal gait recognition from incomplete sequences using rgb-d camera. *IEEE Transactions on Information Forensics and Security (TIFS)*, 9(11):1843–1856, 2014.

[11] Adam Czajka, Daniel Moreira, Kevin Bowyer, and Patrick Flynn. Domain-specific human-inspired binarized statistical image features for iris recognition. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 959–967, 2019.

[12] 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.

[13] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14213–14221, 2020.

[14] 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.

[15] Steven A. Grosz and Anil K. Jain. Latent fingerprint recognition: Fusion of local and global embeddings. *IEEE Transactions on Information Forensics and Security (TIFS)*, 18:5691–5705, 2023.

[16] Bence Gálai and Csaba Benedek. Feature selection for lidar-based gait recognition. In *International Workshop on Computational Intelligence for Multimedia Understanding (IWCIM)*, pages 1–5, 2015.

[17] 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.

[18] Xiao Han, Yiming Ren, Peishan Cong, Yujing Sun Sun, Jingya Wang, Lan Xu, and Yuexin Ma. Gait recognition in large-scale free environment via single liDAR. In *ACM Multimedia 2024*, 2024.

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

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–—1780, nov. 1997.

[22] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[23] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 448–456, Lille, France, 07–09 Jul 2015.

[24] Yumi Iwashita and Ryo Kurazume. Person identification from human walking sequences using affine moment invariants. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 436–441, 2009.

[25] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5010–5019, 2018.

[26] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015.

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

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014.

[29] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[30] Patrick Kozlow, Noor Abid, and Svetlana Yanushkevich. Gait type analysis using dynamic bayesian networks. *Sensors*, 18(10), 2018.

[31] Toby H.W. Lam, K.H. Cheung, and James N.K. Liu. Gait flow image: A silhouette-based gait representation for human identification. *Pattern Recognition*, 44(4):973–987, 2011.

[32] Peter K Larsen, Erik B Simonsen, and Niels Lynnerup. Gait analysis in forensic medicine. *Journal of Forensic Sciences*, 53(5):1149–53, 2008.

[33] Jialian Li, Jingyi Zhang, Zhiyong Wang, Siqi Shen, Chenglu Wen, Yuexin Ma, Lan Xu, Jingyi Yu, and Cheng Wang. Lidarcap: Long-range markerless 3d human motion capture with lidar point clouds. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20470–20480, 2022.

[34] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4089–4098, 2021.

[35] Xiang Li, Yasushi Makihara, Chi Xu, and Yasushi Yagi. End-to-end model-based gait recognition using synchronized multi-view pose constraint. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 4089–4098, 2021.

[36] Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori Hashimoto. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[37] You Li and Javier Ibanez-Guzman. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4):50–61, 2020.

[38] Junhao Liang, Chao Fan, Saihui Hou, Chuanfu Shen, Yongzhen Huang, and Shiqi Yu. Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 375—390, Berlin, Heidelberg, 2022. Springer-Verlag.

[39] B. Lin, S. Zhang, and X. Yu. Gait recognition via effective global-local feature representation and local temporal aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14628–14636, 2021.

[40] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[41] Zhijian Liu, Alexander Amini, Sibo Zhu, Sertac Karaman, Song Han, and Daniela L. Rus. Efficient and robust lidar-based end-to-end navigation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 13247–13254, 2021.

[42] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM Trans. Graph.*, 34(6), October 2015.

[43] 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.

[44] Rodrigo Marcuzzi, Lucas Nunes, Louis Wiesmann, Elias Marks, Jens Behley, and Cyrill Stachniss. Mask4d: End-to-end mask-based 4d panoptic segmentation for lidar sequences. *IEEE Robotics and Automation Letters*, PP:1–8, 11 2023.

[45] Andres Milioto, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Rangenet ++: Fast and accurate lidar semantic segmentation. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4213–4220, 2019.

[46] Daigo Muramatsu, Yasushi Makihara, Haruyuki Iwama, Takuya Tanoue, and Yasushi Yagi. Gait verification system for supporting criminal investigation. In *Proceedings of the IAPR Asian Conference on Pattern Recognition (ACPR)*, pages 747–748, 2013.

[47] Kazuto Nakashima, Yumi Iwashita, and Ryo Kurazume. Generative range imaging for learning scene priors of 3d lidar data. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1256–1266, 2023.

[48] 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.

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

[50] Andrea Sabo, Sina Mehdizadeh, Andrea Iaboni, and Babak Taati. Prediction of parkinsonian gait in older adults with dementia using joint trajectories and gait features from 2d video. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pages 5700–5703, 2021.

[51] Nasrin Sadeghzadehyazdi, Tamal Batabyal, A. Glandon, Nibir K. Dhar, B. O. Familoni, K. M. Iftekharuddin, and Scott T. Acton. Glidar3dj: a view-invariant gait identification via flash lidar data correction. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2606–2610, 2019.

[52] 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.

[53] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 36479–36494, 2022.

[54] 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.

[55] S. Sarkar, P. J. Phillips, Z. Liu, I. R. Vega, P. Grother, and K. W. Bowyer. The humanid gait challenge problem: data sets, performance, and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(2):162–177, 2005.

[56] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015.

[57] Alireza Sepas-Moghaddam and Ali Etemad. Deep gait recognition: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45(1):264–284, 2023.

[58] 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.

[59] Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-kin Wong, and Wang-chun WOO. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.

[60] Makoto Shinzaki, Yumi Iwashita, Ryo Kurazume, and Koichi Ogawara. Gait-based person identification method using shadow biometrics for robustness to changes in the walking direction. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 670–677, 2015.

[61] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *Proceedings of the International Conference on Biometrics (ICB)*, pages 1–8, 2016.

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR.

[63] Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[64] 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.

[65] 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.

[66] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, 2015.

[67] Faezeh Tafazzoli and Reza Safabakhsh. Model-based human gait recognition using leg and arm movements. *Engineering Applications of Artificial Intelligence*, 23(8):1237–1246, 2010.

[68] Yusuke Tashiro, Jiaming Song, Yang Song, and Stefano Ermon. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, pages 24804–24816, 2021.

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

[70] Abhishek Thakur and P Rajalakshmi. L3d-otve: Lidar-based 3-d object tracking and velocity estimation using lidar odometry. *IEEE Sensors Letters*, 8(7):1–4, 2024.

[71] Manoj Tummala and Pravin. A. Lidar sensor for self-driving cars. In *Proceedings of the International Conference on Communication and Electronics Systems (ICCES)*, pages 59–65, 2023.

[72] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

[74] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[75] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[76] Bichen Wu, Alvin Wan, Xiangyu Yue, and Kurt Keutzer. Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 1887–—1893. IEEE Press, 2018.

[77] Bichen Wu, Xuanyu Zhou, Sicheng Zhao, Xiangyu Yue, and Kurt Keutzer. Squeezesegv2: Improved model structure and unsupervised domain adaptation for road-object segmentation from a lidar point cloud. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 4376–—4382. IEEE Press, 2019.

[78] Chenfeng Xu, Bichen Wu, Zining Wang, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. Squeezesegv3: Spatially-adaptive convolution for efficient point-cloud segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 1–—19, 2020.

[79] 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.

[80] Minxiang Ye, Cheng Yang, Vladimir Stankovic Stankovic, Lina Stankovic, and Andrew Kerr. Gait analysis using a single depth camera. In *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 285–289, 2015.

[81] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11779–11788, 2021.

[82] Koki Yoshino, Kazuto Nakashima, Jeongho Ahn, Yumi Iwashita, and Ryo Kurazume. Rgb-based gait recognition with disentangled gait feature swapping. *IEEE Access*, 12:115515–115531, 2024.

[83] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pages 441–444, 2006.

[84] Zhilu Zhang and Mert Sabuncu. Generalized cross entropy loss for training deep neural networks with noisy labels. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[85] Zhonghao Zhang, Yangyang Jiang, Xingyu Cao, Xue Yang, Ce Zhu, Ying Li, and Yipeng Liu. Deep learning based gait analysis for contactless dementia detection system from video camera. In *Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021.

[86] Ziyuan Zhang, Luan Tran, Xi Yin, Yousef Atoum, Xiaoming Liu, Jian Wan, and Nanxin Wang. Gait recognition via disentangled representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4705–4714, 2019.

[87] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20196–20205, 2022.

[88] Vlas Zyrianov, Xiyue Zhu, and Shenlong Wang. Learning to generate realistic lidar point cloud. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 17–—35, 2022.

List of Publication

6.1 Journal Publications

- Learning Viewpoint-Invariant Features for LiDAR-based Gait Recognition. J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume. *IEEE Access*, vol. 11, pp. 129749–129762, 2023.
- RGB-Based Gait Recognition With Disentangled Gait Feature Swapping. K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume. *IEEE Access*, vol. 12, pp. 115515–115531, 2024.

6.2 International Conference

- 2V-Gait: Gait Recognition using 3D LiDAR Robust to Changes in Walking Direction and Measurement Distance. J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pp. 602-607, 2022.
- Gait Recognition using Identity-Aware Adversarial Data Augmentation. K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pp. 596-601, 2022.
- Gait Sequence Upsampling using Diffusion Models for Single LiDAR Sensors. J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pp. 00–00, 2025.
- S2Gait: RGB-based Gait Recognition with Style Feature Sampling Data Augmentation. K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume. In *Proceedings of the IEEE/SICE International Symposium on System Integration (SII)*, pp. 00–00, 2025.

6.3 Domestic Conference (without Review)

- 多層 3D LiDAR と LSTM を用いた距離・点群密度変化に頑健な歩容認証. Jeongho Ahn, 山田弘幸, 中島一斗, 倉爪亮. 第 38 回日本ロボット学会学術講演会, 2D1-05, 2020.10.9-11.
- 3D LiDAR 点群の投影画像を用いた計測距離と歩行方向の変化に頑健な歩容認証. 安正鎬, 中嶋一斗, 吉野弘毅, 岩下友美, 倉爪亮. 第 25 回画像の認識・理解シンポジウム MIRU2022, IS1-93, 2022.7.25-28.
- 歩容特徴の抽出精度向上のための同一人物間の特徴交換を用いた歩容認証. 吉野弘毅, 中嶋一斗, 安正鎬, 岩下友美, 倉爪亮. 第 25 回画像の認識・理解シンポジウム MIRU2022, IS2-77, 2022.7.25-28.
- 3D LiDAR センサの点群投影方式による計測距離と歩行方向に対する歩容認証の頑健性評価. 安正鎬, 中嶋一斗, 吉野弘毅, 岩下友美, 倉爪亮. 第 40 回日本ロボット学会学術講演会 RSJ2022, 4G1-07, 2022.9.5-9.
- 歩容特徴の抽出精度向上のための異なる人物間の特徴交換を用いた歩容認証. 吉野弘毅, 中嶋一斗, 安正鎬, 岩下友美, 倉爪亮. 第 40 回日本ロボット学会学術講演会 RSJ2022, 4G1-08, 2022.9.5-9.
- 複数投影視点の適応的学習に基づく 3D LiDAR を用いた歩容認証. 安正鎬, 中嶋一斗, 吉野弘毅, 岩下友美, 倉爪亮. 第 26 回画像の認識・理解シンポジウム MIRU2023, IS1-99, 2023.7.25-28.
- 歩容特徴に基づく歩行者画像の新規生成を用いた歩容認証. 吉野弘毅, 中嶋一斗, 安正鎬, 岩下友美, 倉爪亮. 第 26 回画像の認識・理解シンポジウム MIRU2023, IS1-97, 2023.7.25-28.
- 拡散モデルを用いた LiDAR 点群投影ベースの歩容映像復元. アン・ジョンホ, 中嶋一斗, 吉野弘毅, 岩下友美, 倉爪亮. 第 27 回画像の認識・理解シンポジウム MIRU2024, IS1-157, 2024.8.6-9.
- RGB 情報を活用したデータ拡張に基づく歩容認証. 吉野弘毅, 中嶋一斗, アン・ジョンホ, 岩下友美, 倉爪亮. 第 27 回画像の認識・理解シンポジウム MIRU2024, IS3-108, 2024.8.6-9.