IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 October 2023, accepted 8 November 2023, date of publication 14 November 2023,
date of current version 22 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3333037

METHODS

Learning Viewpoint-Invariant Features for LiDAR-Based Gait Recognition

JEONGHO AHN[1], (Graduate Student Member, IEEE),
KAZUTO NAKASHIMA[2], (Member, IEEE),
KOKI YOSHINO[1], (Graduate Student Member, IEEE),
YUMI IWASHITA[3], (Senior Member, IEEE), AND RYO KURAZUME[2], (Senior Member, IEEE)
[1]Graduate School of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan
[2]Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka 819-0395, Japan
[3]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91125, USA

Corresponding author: Jeongho Ahn (ahn@irvs.ait.kyushu-u.ac.jp)

This work was supported in part by the Japan Science and Technology Agency (JST) SPRING under Grant JPMJSP2136, and in part by the Japan Society for the Promotion of Science (JSPS) KAKENHI under Grant JP20H00230.

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Ethics Committee of the Graduate School of Information Science and Electrical Engineering at Kyushu University under Application No. 2021-02.

ABSTRACT Gait recognition is a biometric identification method based on individual walking patterns. This modality is applied in a wide range of applications, such as criminal investigations and identification systems, since it can be performed at a long distance and requires no cooperation of interests. In general, cameras are used for gait recognition systems, and previous studies have utilized depth information captured by RGB-D cameras, such as Microsoft Kinect. In recent years, multi-layer LiDAR sensors, which can obtain range images of a target at a range of over 100 m in real time, have attracted significant attention in the field of autonomous mobile robots and self-driving vehicles. Compared with general cameras, LiDAR sensors have rarely been used for biometrics due to the low point cloud densities captured at long distances. In this study, we focus on improving the robustness of gait recognition using LiDAR sensors under confounding conditions, specifically addressing the challenges posed by viewing angles and measurement distances. First, our recognition model employs a two-scale spatial resolution to enhance immunity to varying point cloud densities. In addition, this method learns the gait features from two invariant viewpoints (i.e., left-side and back views) generated by estimating the walking direction. Furthermore, we propose a novel attention block that adaptively recalibrates channel-wise weights to fuse the features from the aforementioned resolutions and viewpoints. Comprehensive experiments conducted on our dataset demonstrate that our model outperforms existing methods, particularly in cross-view, cross-distance challenges, and practical scenarios.

INDEX TERMS Gait recognition, 3D point cloud, LiDAR, convolutional neural networks, attention mechanism.

I. INTRODUCTION

Gait recognition is used to identify people based on their walking patterns. This promising biometric technology has attracted considerable attention because, in this method, distinct physical and behavioral characteristics are used. Compared with other modalities, such as faces, fingerprints, and retinas, gait has the following advantages: it can be easily captured at a long distance and does not require explicit cooperation or contact with the subjects of interest. Camouflaging gait is difficult because of the complex gait dynamics. Therefore, gait recognition exhibits considerable potential for various applications such as criminal investigations and social security.

Although RGB cameras are widely used for capturing gait data because of their low economic cost, such cameras

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

typically make the extraction of gait features difficult because silhouettes, which are commonly used in gait recognition tasks, may not be effective in capturing motion cues, especially in frontal and back views. Studies have focused on this problem by applying RGB-D cameras, such as Microsoft Kinect, which handle additional depth information to effectively use of dynamic textures [1], [2]. Compared with typical RGB cameras, RGB-D cameras have unique advantages, including robustness to various illumination conditions, simple background subtraction, and abundant 3D geometry. However, these RGB-D cameras exhibit severe limitations in measurement distance and field of view, with a maximum of approximately 10 m and 70°, respectively, rendering their application difficult.

LiDAR sensors, which are laser-based range sensors that can measure the surrounding geometry in a 3D point cloud format, have attracted considerable attention in the field of computer vision. These sensors have been applied in mobile robots and self-driving vehicles for computer vision tasks, such as object detection, tracking, and navigation. Compared with RGB-D cameras, LiDAR sensors are more robust to various lighting conditions and can measure longer distances due to their active sensing, which is based on short-wavelength pulsed lasers. Although radar has a long range and is insensitive to lighting fluctuations, LiDAR sensors are more resistant to noise and provide higher spatial resolutions. Therefore, LiDAR-based three-dimensional (3D) perception has become essential for autonomous driving.

Compared with general RGB cameras, LiDAR sensors have rarely been used in biometrics. A possible cause is the lower spatial resolution of LiDAR sensors at long distances, which makes it difficult to capture fine-grained motions of humans in their entirety. Given the benefits of LiDAR, such as robustness to varying illumination conditions, high accuracy in 3D mapping, long-range measurement capabilities beyond those of depth cameras, and a scanning range covering 360° in azimuth, LiDAR can be used for outdoor applications as a biometric identifier. Furthermore, these LiDAR sensors can be used as alternatives to RGB cameras to protect personal information because the direct visual identity-related features of individuals are not extracted.

We previously proposed a gait recognition method using 3D LiDAR [3] and demonstrated the potential of LiDAR sensors for these recognition tasks. In this approach, spatio–temporal features are modeled with depth gait shapes and LSTMs [4]. However, in this study [3], both the measurement distances and walking directions from the sensor were kept constant because the axes of the generated image sequences depended on the specific resolutions constrained by the sensor hardware. This situation is only valid if LiDAR sensors are placed in a corridor or narrow street where people walk in a single direction. In the future, LiDAR sensors will be widely used in numerous scenarios, particularly in mobile robots for person identification. For example, security robots, which can be operated 24 hours a day and are less conspicuous than humans, are becoming increasingly common in malls, offices, and public spaces. Night-time surveillance can be achieved without the requirement for additional vision sensors by applying biometrics to such security robots. Furthermore, self-driving cars can be equipped with biometric devices to detect and identify specific users while driving. Considering these scenarios, designing a robust gait recognition model that accounts for intra-subject changes is critical for maintaining stable recognition capacity.

To minimize the effect of variations irrelevant to gait features, we focused on improving the robustness in two key areas: viewing angles and measurement distances. These conditions are the primary challenges typically encountered in LiDAR-related tasks. The proposed gait recognition model using LiDAR is displayed in Fig. 1. Specifically, we used a two-scale spatial resolution approach to learn various point cloud densities projected onto 2D grids representing depth information. Furthermore, this model exploits gait features from two invariant viewpoints (i.e., left-side and back views) across the gait sequence to enhance the consistency of walking dynamics, whereas these gait shapes cannot be obtained from RGB cameras. In this study, we designed a 2D-attention-based block to fuse gait features from multiple resolutions and viewpoints. Unlike a typical self-attention mechanism that considers the interrelationship of a single input feature, this block takes two different features and compares their statistics to bias towards the more informative one.

A preliminary version of this study was published in [5]. We have extended this version based on the following three aspects: 1) Rather than using pooling approaches [6], [7], we 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) We conducted additional ablation studies to verify the proposed modules. 3) We compared the recognition performance with prior methods on our dataset, which consists of combinations of cross-views and cross-distances, to achieve deeper insights than previous studies [5].

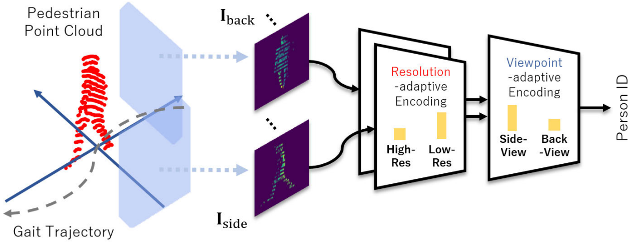


FIGURE 1. Overview of the proposed gait recognition model using 3D LiDAR, which learns two viewpoint-invariant gait shapes in varying point cloud densities using an attention-based approach.

The contributions of this study are as follows:

1) A novel framework is proposed for gait recognition using 3D LiDAR. 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 our dataset revealed that the proposed method surpassed the recognition accuracy of prior studies.

II. RELATED WORK

A. GAIT REPRESENTATION IN AN RGB CAMERA

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.

1) 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 [1], [8]. 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 [9], [10]. In particular, studies [9], [11] have achieved high recognition performance by adopting a 3D human representation [12] 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.

2) 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) [13] 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 [14], gait flow patterns [15], and affine moment invariants [16], have been proposed. Furthermore, the performance is improved by feeding GEIs into CNNs [17], 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. [18] achieved promising results by integrating gait silhouettes as a set [18], and [19] extracted spatio–temporal features from each body part. In contrast, Fan et al. [20] proposed a global–local convolution approach to address the neglect of local region details in gait frames, and a subsequent version [21] 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 [22]. The separation performance was further improved by augmenting the training data through adversarial generation [23]. We 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 [24].

B. 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. [2] classified the gait types of individuals using RGB-D cameras by using a 3D skeleton model with Bayesian networks, whereas Sadeghzadehyazdi et al. [25] applied flash LiDAR sensors with both 2D and 3D skeleton models. In studies applying LiDAR sensors, Benedek et al. [26] proposed a GEI-based method to re-identify individuals in a short time. However, this method cannot satisfactorily extract dynamic features from gait frames, which are critical for discriminating individuals. To address this problem, Yamada et al. [3] proposed a method for exploring temporal gait changes using LSTMs [4]. They exploited depth representation in a spherical projection, which has been used in several LiDAR-related tasks for its processing efficiency [27]. 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.

C. 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 [28]. Several studies have demonstrated its applicability to computer vision. For example, Wang et al. [29] proposed a nonlocal operation that captures long-range dependencies in images or videos and can be plugged into CNN-based architectures. Hu et al. [30] enhanced the representation power of CNNs by focusing on channel relationships using a gating mechanism, whereas Woo et al. [31] 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.

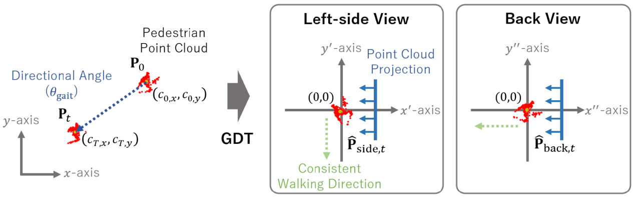


FIGURE 2. Overview of the gait direction transformation (GDT) process that generates two invariant gait shapes from pedestrian point cloud sequences.

III. METHODOLOGY

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

A. GAIT DIRECTION TRANSFORMATION

We first describe the gait direction transformation (GDT) process for estimating the walking directions of point cloud sequences. These 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. 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, we define the center of mass $\mathbf{c}_t = (c_{t,x}, c_{t,y}, c_{t,z})$ for the time step t as follows:

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

where $c_{t,z}$ is set to zero because we only consider the walking direction on the xy-plane. 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:

$$\theta_{\text{gait}} = \arctan(c_{T,y} - c_{0,y}, c_{T,x} - c_{0,x}), \tag{2}$$

where $\arctan(\cdot,\cdot)$ calculates the angle between a pedestrian and a positive x-axis in the Euclidean plane. 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:

$$\hat{\mathbf{p}}_{t,n} = \mathbf{R}_z(-\theta_{\text{gait}} - \frac{\pi}{2}) \cdot (\mathbf{p}_{t,n} - \mathbf{c}_t), \tag{3}$$

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. 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 $\mathbf{R}_z(\cdot)$.

B. DEPTH IMAGE GENERATION

We generated input data 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. Our 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. Our projection manner differs from the approach used in [3], in which point clouds are assigned to an angular grid using spherical projection, as displayed in Fig. 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 [3], 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:

$$v = \left\lfloor \frac{1}{R_z} \cdot (\hat{p}_{t,n,z} - \min_{n\in\{1,\ldots,N\}}(\hat{p}_{t=0,n,z}) + l_{z-\text{const}}) \right\rfloor, \tag{4}$$

$$h = \left\lfloor \frac{1}{R_y} \cdot (\hat{p}_{t,n,y} + \frac{l_y}{2}) \right\rfloor, \tag{5}$$

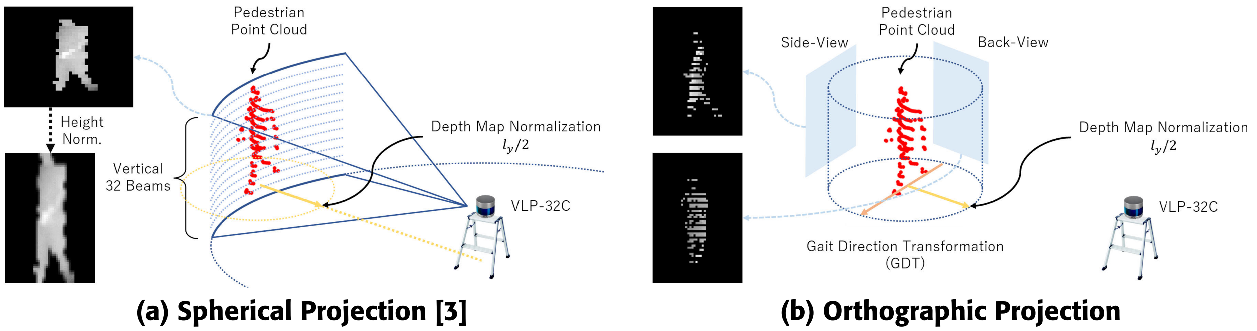


FIGURE 3. Comparison between the prior spherical and proposed orthographic projection approaches, representing gait shapes with depth information.

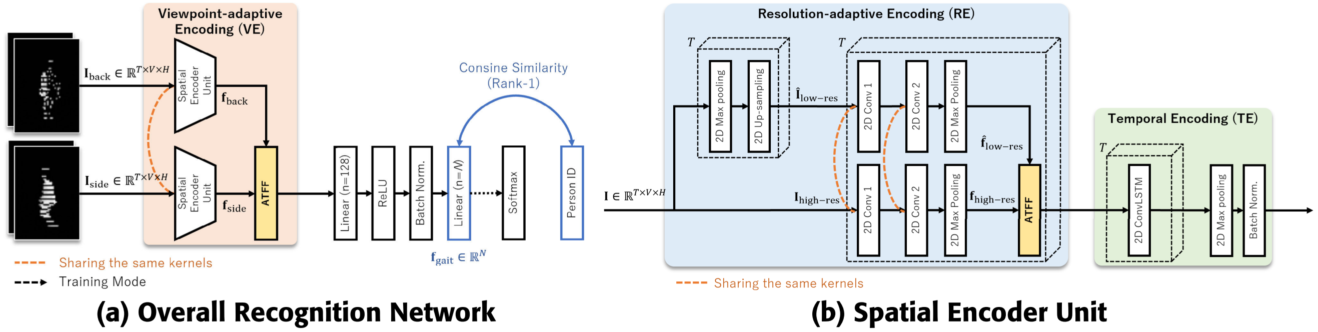

Viewpoint-adaptive
Encoding (VE)
$\mathbf{I}_{\text{back}} \in \mathbb{R}^{T\times V\times H}$
$\mathbf{I}_{\text{side}} \in \mathbb{R}^{T\times V\times H}$
Spatial
Encoder
Unit
$\mathbf{f}_{\text{back}}$
$\mathbf{f}_{\text{side}}$
ATFF
Linear (n=128)
ReLU
Batch Norm.
Linear (n=N)
Softmax
Person ID
Consine Similarity
(Rank-1)
$\mathbf{f}_{\text{gait}} \in \mathbb{R}^N$
Sharing the same kernels
Training Mode
(a) Overall Recognition Network
Resolution-adaptive Encoding (RE)
T
2D Max pooling
2D Up-sampling
$\hat{\mathbf{I}}_{\text{low-res}}$
$\mathbf{I} \in \mathbb{R}^{T\times V\times H}$
$\mathbf{I}_{\text{high-res}}$
2D Conv 1
2D Conv 2
2D Max Pooling
$\hat{\mathbf{f}}_{\text{low-res}}$
$\mathbf{f}_{\text{high-res}}$
ATFF
Temporal Encoding (TE)
2D ConvLSTM
2D Max pooling
Batch Norm.
Sharing the same kernels
(b) Spatial Encoder Unit

FIGURE 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.

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, we can obtain the gait image sequence $\mathbf{I} = (\mathbf{i}_1, \ldots, \mathbf{i}_T) \in \mathbb{R}^{T\times V\times H}$ for T frames. 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.

C. RECOGNITION NETWORK

In this section, we describe a recognition network for learning the discriminative gait features from the aforementioned inputs. The architecture in Fig. 4 (a) 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. 4 (b). 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.

1) 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 pervious version of this study [5], 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, we 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, we use the nearest neighbor algorithm (i.e., rank-1 accuracy) to compute the cosine similarity between galleries and probes for subsequent evaluations.

2) 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. 5.

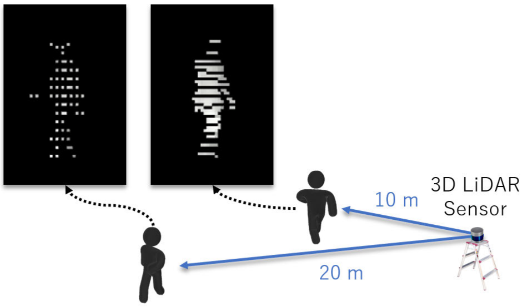


FIGURE 5. Examples with various measurement distances, with different sparsity of proposed depth images.

Gait-related spatial features 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, we 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:

$$\mathbf{I}_{\text{low-res}} = \text{Maxpool2D}(\mathbf{I}) \tag{6}$$

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

$$\hat{\mathbf{I}}_{\text{low-res}} = \text{Upsampling2D}(\mathbf{I}_{\text{low-res}}) \tag{7}$$

Subsequently, we feed two gait image sequences 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 1.

TABLE 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	-

3) ATTENTION-BASED TWO FEATURES FUSING

As displayed in Fig. 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. 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 [30], we 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 $\mathbf{f}_2 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$, we first compress the global spatial information 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 [30], our 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:

$$z_{1,c_{\text{attn}}} = \frac{1}{T\times H_{\text{attn}}\times W_{\text{attn}}}\sum_{i=1}^{T}\sum_{j=1}^{H_{\text{attn}}}\sum_{k=1}^{W_{\text{attn}}} f_{1,c_{\text{attn}}}(i,j,k). \tag{8}$$

Subsequently, we follow this operation with a second operation that fully captures channel-wise dependencies and expresses probabilistic values to determine which of the two

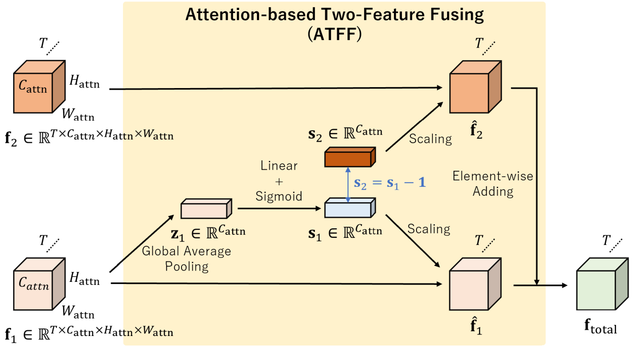

Attention-based Two-Feature Fusing
(ATFF)
T
C_{attn}
H_{attn}
W_{attn}
$\mathbf{f}_2 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$
$\mathbf{s}_2 \in \mathbb{R}^{C_{\text{attn}}}$
Scaling
$\hat{\mathbf{f}}_2$
Linear
+
Sigmoid
$\mathbf{s}_2 = \mathbf{s}_1 - \mathbf{1}$
Element-wise
Adding
$\mathbf{z}_1 \in \mathbb{R}^{C_{\text{attn}}}$
Global Average
Pooling
$\mathbf{s}_1 \in \mathbb{R}^{C_{\text{attn}}}$
Scaling
T
C_{attn}
H_{attn}
W_{attn}
$\mathbf{f}_1 \in \mathbb{R}^{T\times C_{\text{attn}}\times H_{\text{attn}}\times W_{\text{attn}}}$
$\hat{\mathbf{f}}_1$
$\mathbf{f}_{\text{total}}$

FIGURE 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.

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:

$$\mathbf{s}_2 = \mathbf{1} - \mathbf{s}_1 = \mathbf{1} - \sigma(\mathbf{W}_2\delta(\mathbf{W}_1\mathbf{z}_1)), \tag{9}$$

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, we ensure that 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:

$$\mathbf{f}_{\text{total}} = \hat{\mathbf{f}}_1 \oplus \hat{\mathbf{f}}_2 = (\mathbf{s}_1 * \mathbf{f}_1) \oplus (\mathbf{s}_2 * \mathbf{f}_2), \tag{10}$$

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.

4) 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 [32], which is an extension of LSTMs [4] and is used for modeling the spatio–temporal representation of gait features, equipped with 2D max pooling and batch normalization layers [33]. The hyperparameters of this layer are listed in Table 1. Compared with [3], 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.

IV. EXPERIMENTS

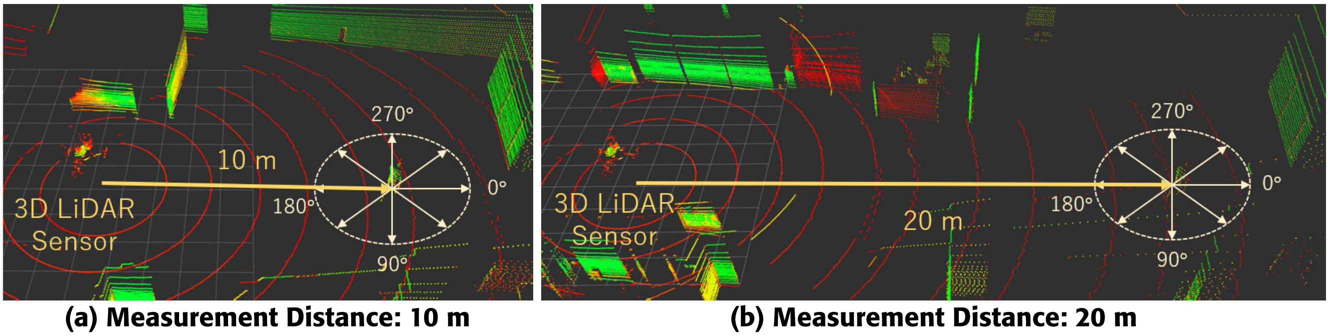
In this section, we report the results of comprehensive experiments performed to evaluate the performance of the proposed method. First, we describe our dataset used in these experiments and compare our method with existing methods under various settings, including measurement distances and viewing angles from LiDAR sensors. We conducted ablation studies to evaluate the effectiveness of proposed modules. Furthermore, we investigated the practicality by limiting the viewing angles of the database. Finally, we visualized the gait features by reducing their dimensions and performed qualitative evaluations.

A. DATASET

To verify the robustness to changes in viewing angles and distances measured from a sensor, we collected a gait database using a single Velodyne VLP-32C, which creates 3D range images with a horizontal 360° field of view using 32 lasers for vertical resolution. 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 the capture, we placed the LiDAR sensor on a tripod at a height of 1.2 m. We 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. 7. We 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. In this experiment section, the viewing angles are determined by the pedestrian's walking direction relative to the sensor-pedestrian vector, as shown in Fig. 7. Compared with other gait datasets [34], [35] commonly used in gait recognition challenges, this combination not only includes cross-view but also cross-distance conditions. Furthermore, 126-point cloud sequences were obtained for each subject under a single condition. Therefore, our gait dataset contains 30 × 8 × 2 × 126 = 60,480 sequences. During training and evaluation, we only used the pedestrian point cloud sequences, which were extracted through background subtraction processing.

B. IMPLEMENTATION DETAILS

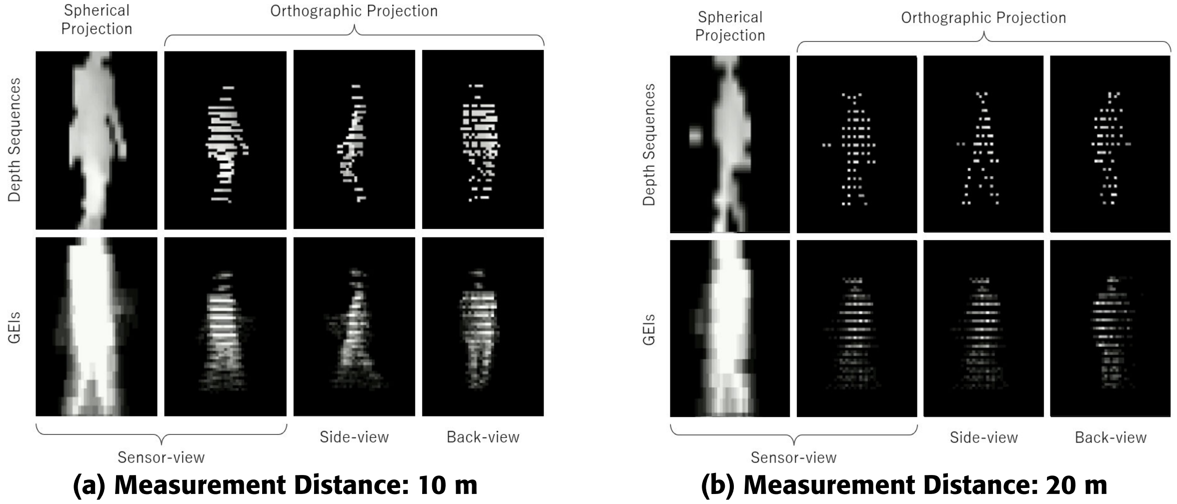
We conducted experiments based on our dataset. Essentially, the first 20 subjects were used for training, and the remaining 10 subjects were used for testing with no overlap. Thus, the training set contains 20 × 8 × 2 × 126 = 40,320 sequences for all experimental settings. In addition, we standardized 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. 8. For optimization, we used the cross-entropy loss to train the networks. We adopted ADAM [36] 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 with Pytorch



(a) Measurement Distance: 10 m

(b) Measurement Distance: 20 m

FIGURE 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 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.

1.12.1, and performed on a single NVIDIA GeForce RTX 3090 GPU. During testing, we used the penultimate feature to match the 10 subjects that were not used in the training.

C. COMPARED METHODS

To evaluate the contribution of each component to the overall performance, we investigated the effectiveness of the proposed method by examining it from three changeable perspectives: viewpoint, point cloud projection, and recognition network. In the GDT, we investigated the extent to which viewing angle-independent gait features, which cannot be directly obtained from RGB cameras, contribute to individual recognition. Furthermore, we evaluated the proposed point cloud projection for input and compared it with the existing method (i.e., spherical projection) used in [3]. When applying the spherical projection, we normalized gait images to a size of 64 × 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, we compared the performance of the proposed network with three prior approaches [3], [17], [26]. Among these approaches, the LGEI-based technique [26] 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, [17], called GEINet, is the most representative network that uses GEIs in the gait recognition field, which structure is similar to that of [26]. In [3], gait image sequences with depth information are input into CNN and LSTM layers to classify individuals. We compared our approach with both Network 1 and 2 in [3], where the second architecture is a modified version of the first method by supplementing a subtraction operation at the front of the the LSTMs.

D. MAIN RESULTS

Table 2 presents a comparison between the proposed method and the prior approaches. The results for all networks were obtained through experiments. In this experiment, we 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

TABLE 2. Comparison with prior studies on our 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. [26]	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. [17]	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) [3]	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) [3]	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. [5]	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
		Ours	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. [26]	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. [17]	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) [3]	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) [3]	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. [5]	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
		Ours	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**

the typical cross-view conditions. In this experiment, the gallery and probe contained 10 × 1 × 7 × 42 = 2,940, and 10 × 1 × 1 × 84 = 840 sequences for each condition, respectively. In the proposed network with a single viewpoint, we used only one spatial encoder unit in the VE module without the ATFF block.

First, in Table 2, we observed that networks using our 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. We achieved improved recognition performance by using both the GDT and our proposed projection (Ortho.) except for Network 1 [3]. 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, we observed a performance decline in the networks that use depth image sequences [3], [5]. This result indicates that the spherical

TABLE 3. Comparison with prior studies on our dataset under the sole cross-view condition (%).

Networks	Modalities	Projections	Viewpoints			Mean
			Sensor-view	Side-view	Back-view	
Benedek et al. [26]	GEI	Spher.	✓			62.6
				✓		71.7
		Ortho.	✓			71.9
				✓		76.5
Shiraga et al. [17]	GEI	Spher.	✓			70.9
				✓		71.1
		Ortho.	✓			71.9
				✓		76.5
Yamada et al. (Network 1) [3]	Depth Seq.	Spher.	✓			53.4
				✓		49.8
		Ortho.	✓			64.1
				✓		62.1
Yamada et al. (Network 2) [3]	Depth Seq.	Spher.	✓			44.1
				✓		54.0
		Ortho.	✓			54.9
				✓		61.2
Ahn et al. [5]	Depth Seq.	Spher.		✓	✓	93.4
		Ortho.		✓	✓	**94.8**
Ours	Depth Seq.	Spher.	✓			89.6
				✓		54.6
					✓	83.4
				✓	✓	89.9
		Ortho.	✓			89.7
				✓		86.1
					✓	91.0
				✓	✓	93.8

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 [3] because they negatively affect the depth values in gait images processed with the GDT. In our proposed network and the previous version [5], 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, we observed that our fixed viewpoint strategy positively affected GEI-based networks [17], [26] 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, our complete method achieved the highest discriminative capability by using two viewpoints.

Table 3 presents the performance results for the recognition networks under the sole cross-view condition. This experimental setting is identical to that in Table 2, except that the measurement distances for both the gallery and probe are the same. Each average value in Table 3 represents a combination of all eight cross-views and two distances. We observed that the results in Table 3 are generally higher than those in Table 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 [5]. From these results, we observed that while the pooling method that combines two gait features from independently trained units may achieve optimal performance under a cross-view condition, our proposed attention method is more effective in scenarios where the point cloud density changes.

E. ABLATION STUDY

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

TABLE 4. Effect of input modalities and temporal aggregating manners (%).

Modalities		Temporal Encoding (TE)		Mean
Silhouette Seq.	Depth Seq.	1D-LSTM	ConvLSTM [32]	
✓		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

1) MODALITY AND TE

We first investigated the effectiveness of our network by changing the input modalities, temporal aggregating manners, and hyper-parameters, as presented in Table 4. In this case, we 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, we enhanced the recognition performance by replacing the previously used LSTMs in [3] with ConvLSTMs [32] to effectively extract spatio–temporal features. The accuracies shown in Table 4 represent the averages across all the cross-view and cross-distance combinations, as in Sec. IV-D. 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.

2) IMPACT OF RE

In Table 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 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

TABLE 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 [30]			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

in Fig. 4 (b) and Fig. 6. The average accuracies obtained using a single resolution were lower than the results in the last four rows of Table 5, using only a single experiment by changing the components of the VE module 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 [30]. 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 [30] that considers channel-wise interrelationships, in our 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 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.

TABLE 6. Ablation experiment for viewpoint-adaptive encoding (VE) (%).

Original view	Side-view	Back-view	Fusion	Mean
✓				72.1
	✓			73.4
		✓		77.3
	✓	✓	Average Pooling [5]	79.1
	✓	✓	Max Pooling	78.5
	✓	✓	Concatenating	77.3
	✓	✓	ATFF ($T = 1$)	**81.2**

3) IMPACT OF VE

Table 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 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 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 [5], 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.

F. 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, we investigated the practicality of the proposed model, comparing with the prior methods [3], [17], [26]. The experiment was conducted by restricting the viewing angle of the galleries. Specifically, we 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. We evaluated the recognition models from three perspectives, namely projection, viewpoint, and network, for which the combinations are the same as shown in Sec. IV-D. Each accuracy value in Table 7 is the average of eight probe views and two cross-distances per viewing angle of the gallery.

In Table 7, our 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 [3]. 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 our attention manner, achieved the best performance for all combinations, from all viewing angles of the galleries.

TABLE 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. [26]	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. [17]	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) [3]	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) [3]	Depth Seq.	Spher.	✓			31.0	28.2	33.6
				✓		15.2	15.8	17.3
		Ortho.	✓			33.5	41.9	45.8
				✓		43.4	46.6	43.4
Ours	Depth Seq.	Spher.	✓			39.1	53.4	39.5
				✓		50.8	47.5	48.3
					✓	40.4	49.6	47.0
				✓	✓	50.9	49.5	52.1
		Ortho.	✓			64.3	62.4	68.9
				✓		67.8	61.3	66.6
					✓	63.3	67.7	67.4
				✓	✓	**73.0**	**70.2**	**72.7**

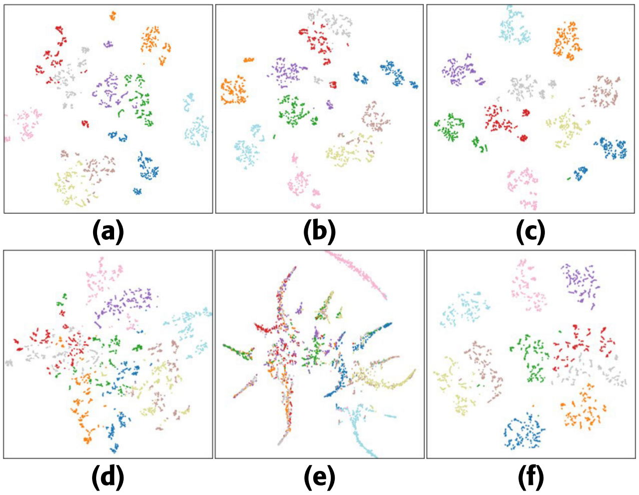


FIGURE 9. t-SNE visualization of the gait features from 10 subjects, each with 8 views, 2 distances, and 42 sequences. The top row shows the proposed model, in which the RE and VE modules are applied in order from left to right. In this case, (a), (b), and (c) correspond to the top line of Table 5 and the top and bottom lines of Table 6, respectively. The bottom row shows the prior methods, with [3], [5], and [17] 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.

G. FEATURE VISUALIZATION THROUGH T-SNE

In this section, we describe a qualitative evaluation by applying t-SNE [37] to visualize gait features through a 2D manifold space. Using the learned recognition models, we 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. 9 presents a visualization of gait features extracted from our 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. 9. We visualized the features of the prior methods in the bottom row of Fig. 9, representing the results of [3], [17], and [5] 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 Sec IV-D. The distances of points between intra- and inter-class reveal that our complete model exhibits superior discrimination power, compared with the three prior methods.

V. CONCLUSION

We proposed a depth-based gait recognition model using 3D LiDAR, which is robust to changes in viewing angles and distances captured from sensors. Focusing on the two conditions, we enhanced the discriminative ability from three perspectives: point cloud projection, GDT, and recognition network. Specifically, in the proposed method, the gait shapes of two invariant viewpoints (i.e., side and back views) are generated from the point cloud sequence, and gait features are extracted from them using a novel attention method that fuses two similar features effectively. Experiments on our dataset indicated that the proposed approach achieved the best recognition performance in both cross-view and cross-distance challenges compared with prior approaches. Furthermore, based on the results of conducted extensive experiments, the proposed model exhibited considerable potential for use in practical scenarios, such as few viewing angles.

REFERENCES

[1] M. Balazia and P. Sojka, ''You are how you walk: Uncooperative MoCap gait identification for video surveillance with incomplete and noisy data,'' in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2017, pp. 208–215.

[2] P. Kozlow, N. Abid, and S. Yanushkevich, ''Gait type analysis using dynamic Bayesian networks,'' *Sensors*, vol. 18, no. 10, p. 3329, Oct. 2018.

[3] H. Yamada, J. Ahn, O. M. Mozos, Y. Iwashita, and R. Kurazume, ''Gait-based person identification using 3D LiDAR and long short-term memory deep networks,'' *Adv. Robot.*, vol. 34, no. 18, pp. 1201–1211, Sep. 2020.

[4] S. Hochreiter and J. Schmidhuber, ''Long short-term memory,'' *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[5] J. Ahn, K. Nakashima, K. Yoshino, Y. Iwashita, and R. Kurazume, ''2 V-gait: Gait recognition using 3D LiDAR robust to changes in walking direction and measurement distance,'' in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 602–607.

[6] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, ''Multi-view convolutional neural networks for 3D shape recognition,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 945–953.

[7] A. Kanezaki, Y. Matsushita, and Y. Nishida, ''RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5010–5019.

[8] F. Tafazzoli and R. Safabakhsh, ''Model-based human gait recognition using leg and arm movements,'' *Eng. Appl. Artif. Intell.*, vol. 23, no. 8, pp. 1237–1246, Dec. 2010.

[9] X. Li, Y. Makihara, C. Xu, Y. Yagi, S. Yu, and M. Ren, ''End-to-end model-based gait recognition,'' in *Proc. Asian Conf. Comput. Vis. (ACCV)*, 2020, pp. 3–20.

[10] X. Li, Y. Makihara, C. Xu, and Y. Yagi, ''End-to-end model-based gait recognition using synchronized multi-view pose constraint,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 4106–4115.

[11] J. Zheng, X. Liu, W. Liu, L. He, C. Yan, and T. Mei, ''Gait recognition in the wild with dense 3D representations and a benchmark,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20196–20205.

[12] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, ''SMPL: A skinned multi-person linear model,'' *ACM Trans. Graph.*, vol. 34, no. 6, p. 248, Oct. 2015.

[13] J. Han and B. Bhanu, ''Individual recognition using gait energy image,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006.

[14] M. Shinzaki, Y. Iwashita, R. Kurazume, and K. Ogawara, ''Gait-based person identification method using shadow biometrics for robustness to changes in the walking direction,'' in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Jan. 2015, pp. 670–677.

[15] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, ''Gait flow image: A silhouette-based gait representation for human identification,'' *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, Apr. 2011.

[16] Y. Iwashita and R. Kurazume, ''Person identification from human walking sequences using affine moment invariants,'' in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2009, pp. 436–441.

[17] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, ''GEINet: view-invariant gait recognition using a convolutional neural network,'' in *Proc. Int. Conf. Biometrics (ICB)*, Jun. 2016, pp. 1–8.

[18] H. Chao, K. Wang, Y. He, J. Zhang, and J. Feng, ''GaitSet: Cross-view gait recognition through utilizing gait as a deep set,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3467–3478, Jul. 2022.

[19] C. Fan, Y. Peng, C. Cao, X. Liu, S. Hou, J. Chi, Y. Huang, Q. Li, and Z. He, ''GaitPart: Temporal part-based model for gait recognition,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14213–14221.

[20] B. Lin, S. Zhang, and X. Yu, ''Gait recognition via effective global–local feature representation and local temporal aggregation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 14628–14636.

[21] J. Liang, C. Fan, S. Hou, C. Shen, Y. Huang, and S. Yu, ''GaitEdge: Beyond plain end-to-end gait recognition for better practicality,'' 2022, *arXiv:2203.03972*.

[22] Z. Zhang, L. Tran, F. Liu, and X. Liu, ''On learning disentangled representations for gait recognition,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 345–360, Jan. 2022.

[23] K. Yoshino, K. Nakashima, J. Ahn, Y. Iwashita, and R. Kurazume, ''Gait recognition using identity-aware adversarial data augmentation,'' in *Proc. IEEE/SICE Int. Symp. Syst. Integr. (SII)*, Jan. 2022, pp. 596–601.

[24] J. Li, J. Zhang, Z. Wang, S. Shen, C. Wen, Y. Ma, L. Xu, J. Yu, and C. Wang, ''LiDARCap: Long-range markerless 3D human motion capture with LiDAR point clouds,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 20470–20480.

[25] N. Sadeghzadehyazdi, T. Batabyal, A. Glandon, N. K. Dhar, B. O. Familoni, K. M. Iftekharuddin, and S. T. Acton, ''GLiDAr3DJ: A view-invariant gait identification via flash LiDAR data correction,'' 2019, *arXiv:1905.00943*.

[26] C. Benedek, B. Gálai, B. Nagy, and Z. Jankó, ''LiDAR-based gait analysis and activity recognition in a 4D surveillance system,'' *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 1, pp. 101–113, Jan. 2018.

[27] K. Nakashima, Y. Iwashita, and R. Kurazume, ''Generative range imaging for learning scene priors of 3D LiDAR data,'' in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 1256–1266.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2017, p. 6000—6010.

[29] X. Wang, R. Girshick, A. Gupta, and K. He, ''Non-local neural networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[30] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[31] S. Woo, J. Park, J. Lee, and I. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, p. 3—19.

[32] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. Wong, and W. Woo, ''Convolutional LSTM network: A machine learning approach for precipitation nowcasting,'' in *Proc. Neural Inf. Process. Syst. (NIPS)*, 2015, p. 802.

[33] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' 2015, *arXiv:1502.03167*.

[34] S. Yu, D. Tan, and T. Tan, ''A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,'' in *Proc. Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 441–444.

[35] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 162–177, Feb. 2005.

[36] D. P. Kingma and J. Ba, ''Adam: A method for stochastic optimization,'' 2014, *arXiv:1412.6980*.

[37] L. van der Maaten and G. Hinton, ''Visualizing data using t-SNE,'' *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

JEONGHO AHN (Graduate Student Member, IEEE) received the B.E. degree in electronic engineering from Gachon University, South Korea, in 2019, and the M.E. degree from Kyushu University, Japan, in 2021, where he is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Electrical Engineering. His research interests include computer vision, machine learning, and biometrics.

KAZUTO NAKASHIMA (Member, IEEE) received the M.Eng. degree, in 2017, and the Ph.D. degree, in 2020. He is an Assistant Professor with the Faculty of Information Science and Electrical Engineering, Kyushu University, Japan. From 2019 to 2020, he was a Research Fellow with the Japan Society for the Promotion of Science (JSPS). His research focuses on computer vision for robotics applications.

KOKI YOSHINO (Graduate Student Member, IEEE) received the B.E. and M.E. degrees from Kyushu University, Japan, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree with the Graduate School of Information Science and Electrical Engineering. His current research interests include machine learning, computer vision, and biometrics.

YUMI IWASHITA (Senior Member, IEEE) received the Ph.D. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, Japan. She is a Robotics Scientist with the Aerial and Orbital Image Analysis Group, NASA's Jet Propulsion Laboratory (JPL). Prior to joining JPL, she was an Associate Professor with Kyushu University. Her research focuses on computer vision for robotics and intelligence, surveillance, and reconnaissance (ISR) applications.

RYO KURAZUME (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees in mechanical engineering from the Tokyo Institute of Technology, in 1989 and 1998, respectively. He is a Professor with the Graduate School of Information Science and Electrical Engineering, Kyushu University. He was the Director of the Robotics Society of Japan (RSJ), from 2009 to 2011 and 2014 to 2015; the Society of Instrument and Control Engineers (SICE), from 2013 to 2015; and the Japan Society of Mechanical Engineers (JSME), from 2021 to 2023. He was the Chairperson of the JSME Robotics and Mechatronics Division, in 2019. He received the JSME Robotics and Mechatronics Academic Achievement Award, in 2012; the RSJ Fellow, in 2016; the SICE System Integration Division Academic Achievement Award, in 2017; the JSME Fellow, in 2018; the SICE Fellow, in 2019; and the JSME Robotics and Mechatronics Division Robotics and Mechatronics Award, in 2021. His current research interests include legged robot control, computer vision, multiple mobile robots, service robots, care technology, and biometrics.