Camera-LiDAR Cross-modality Gait Recognition

Wenxuan Guo[1*], Yingping Liang[2*], Zhiyu Pan[1], Ziheng Xi[1], Jianjiang Feng[1†], and Jie Zhou[1]

[1] Department of Automation, BNRist, Tsinghua University
[2] Beijing Institute of Technology
{gwx22, pzy20, xizh21}@mails.tsinghua.edu.cn liangyingping@bit.edu.cn
{jfeng, jzhou}@tsinghua.edu.cn

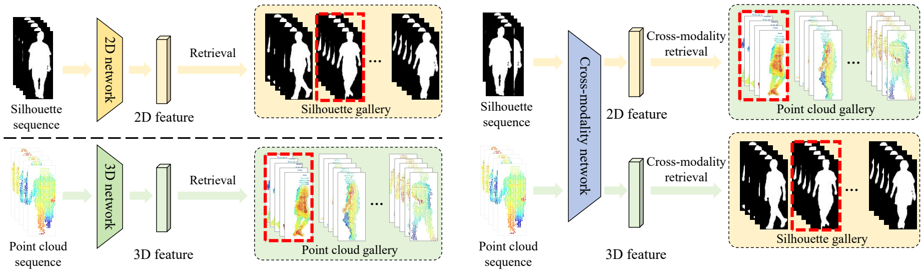
Abstract. Gait recognition is a crucial biometric identification technique. Camera-based gait recognition has been widely applied in both research and industrial fields. LiDAR-based gait recognition has also begun to evolve most recently, due to the provision of 3D structural information. However, in certain applications, cameras fail to recognize persons, such as in low-light environments and long-distance recognition scenarios, where LiDARs work well. On the other hand, the deployment cost and complexity of LiDAR systems limit its wider application. Therefore, it is essential to consider cross-modality gait recognition between cameras and LiDARs for a broader range of applications. In this work, we propose the first cross-modality gait recognition framework between **C**amera and **L**iDAR, namely **CL-Gait**. It employs a two-stream network for feature embedding of both modalities. This poses a challenging recognition task due to the inherent matching between 3D and 2D data, exhibiting significant modality discrepancy. To align the feature spaces of the two modalities, i.e., camera silhouettes and LiDAR points, we propose a contrastive pre-training strategy to mitigate modality discrepancy. To make up for the absence of paired camera-LiDAR data for pre-training, we also introduce a strategy for generating data on a large scale. This strategy utilizes monocular depth estimated from single RGB images and virtual cameras to generate pseudo point clouds for contrastive pre-training. Extensive experiments show that the cross-modality gait recognition is very challenging but still contains potential and feasibility with our proposed model and pre-training strategy. To the best of our knowledge, this is the first work to address cross-modality gait recognition. The code and dataset are available at https://github.com/GWxuan/CL-Gait.

Keywords: Gait recognition · Cross-modality · Contrastive pre-training

1 Introduction

Gait recognition is a crucial long-range identification technology without physical contact, which has great advantages in privacy protection and cross-clothing

[*] Equal contribution. [†] Corresponding author.



(a) Camera-based gait recognition (top), and LiDAR-based gait recognition (bottom).

(b) Our proposed cross-modality gait recognition between camera and LiDAR.

Fig. 1: Overview of single-modality and cross-modality gait recognition. Single-modality gait recognition takes data of one modality as input and searches within a gallery of the same data type. In contrast, cross-modality gait recognition processes two modalities and identify individuals within a gallery of a different modality.

recognition [2]. It identifies individuals based on behavioral characteristics extracted from walking sequences recorded by sensors. Gait recognition has been extensively applied in fields such as user identification, sports science, and public security. Most current gait recognition research employs cameras as sensors [29]. With the advancement of computer vision technology, camera-based gait recognition continues to achieve new progress [16–19, 22, 32].

While utilizing cameras for gait recognition is cost-effective, it faces limitations in certain applications, such as low-light environments, long-distance recognition, and scenarios requiring precise 3D perception. In low-light conditions, the available information in RGB images significantly diminishes [9], making it challenging to accurately detect pedestrians and segment their silhouettes. Furthermore, due to the decline in image clarity, camera-based systems struggle to distinguish the fine-grained features from a distance [5]. Additionally, accurately capturing the 3D movement of individuals, crucial for comprehensive gait analysis, is beyond the capability of traditional camera systems [30].

Recently, to address the challenges under low light environment faced by RGB cameras, some researches [21, 23, 25, 40] have focused on RGB-Infrared cross-modality person re-identification (ReID), achieving commendable performance. However, infrared cameras struggle to capture pedestrians wearing insulating or thick clothing. Besides, they are constrained by limited detection distance and resolution, as well as susceptibility to lighting conditions and weather. It also faces the risk of pedestrian privacy and safety, compared with gait recognition. These limitations highlight the need for exploring alternative or complementary technologies that can offer more robust solutions for person identification tasks across a variety of environmental conditions.

In recent years, LiDAR technology has been increasingly utilized [35] due to its impressive gains on 3D information. Beyond its mature applications in autonomous driving, LiDAR has also been applied in individual identification areas, such as gait recognition [30] and person re-identification [9] most recently.

Through point cloud data, LiDAR can provide precise 3D perception, aiding in the acquisition of an individual's 3D geometric structure. From this, intrinsic characteristics of a person, including height, body shape, and gait, can be extracted. Moreover, LiDARs possess long-distance perception capabilities, unaffected by lighting conditions and complex backgrounds. And it offers enhanced privacy protection than camera [9]. The robustness to environmental variables and the ability to capture detailed biometric data make it an invaluable tool for advanced surveillance and security systems. However, compared to camera systems, LiDAR's deployment complexity and cost are somehow higher, making it unsuitable for large-scale deployment in general scenarios.

Considering the respective advantages of cameras and LiDARs, switching sensors in different scenarios to explore cross-modality gait recognition presents a worthwhile task for investigation. Cross-modality gait recognition offers an option with a broader application scope. Cameras can be utilized in general scenes with normal lighting, while LiDAR can be applied in specific scenarios, offering a complementary tool that enhances overall recognition capabilities.

In this work, we present the first study on camera-LiDAR cross-modality gait recognition as shown in Fig. 1. This task poses significant challenges due to the need for matching data between 3D and 2D modalities. Based on our observations, there are substantial differences between point clouds and images. To be specific, for gait recognition, point clouds primarily focus on the 3D positioning of body parts, whereas images capture precise contour information of individuals. To address these challenges and explore the feasibility of cross-modality gait recognition, we propose a novel cross-modality framework, namely CL-Gait.

To extract gait features from both modalities, CL-Gait basically employs a two-stream network. Considering modality differences and data processing consistency, we first project and upsample the point clouds to obtain depth images with the same resolution as silhouettes, and use the projected depth images to train the network. Furthermore, to mitigate modality discrepancy, CL-Gait utilizes a contrastive silhouette-point pre-training approach (CSPP) to align the feature spaces of the two modalities. Pre-training requires aligned camera-LiDAR data under the same view, which is hard to obtain. To this end, we propose a new strategy of generating cross-modality gait data for pre-training. This strategy utilizes advances in monocular depth estimation and is able to generate large-scale dataset for pre-training using only single RGB images.

Extensive experiments have revealed the following **three** insights: **1)** CL-Gait achieves an average rank-1 accuracy of 54.21%. The results show at least 22.90% improvement over the baseline models and demonstrate the significant potential of cross-modality gait recognition. **2)** Utilizing large-scale depth images generated from point clouds as input is proved superior to using point clouds directly. **3)** The strategy of contrastive pre-training on large-scale generated paired data mitigates the modality differences, contributing to performance improvement.

To summarize, our main contributions are as follows:

- To the best of our knowledge, this is the first work on cross-modality gait recognition. Extensive experiments show the potential of utilizing camera and LiDAR for gait recognition in challenging cross-modality scenarios.
- We analyse several network structures and conduct comparable experiments on their effectiveness for cross-modality gait recognition. Based on the two-stream network with better performance, our proposed CL-Gait is capable of matching camera silhouettes and LiDAR points.
- We propose a contrastive pre-training method to align the feature spaces of the two modalities. To make up for the absence of paired camera-LiDAR data, we further introduce a large-scale data generation strategy.

2 Related Work

Gait Recognition. Based on the sensor used, exiting gait recognition methods can be categorized into camera-based and LiDAR-based methods.

Camera-based gait recognition has been extensively studied over the past decade. The majority of camera-based methods focus on extracting appearance features directly from images or videos [2, 6], adapting well to resolution reduction and achieving impressive performance [5]. These methods utilize silhouettes [16, 18, 19] or other gait templates [1, 34] for spatial feature extraction and temporal modeling. Additionally, some researchers have explored using estimated underlying structure of human body [15, 17, 32], such as 2D/3D pose and the SMPL model [22], as inputs. While theoretically robust against factors like carrying and clothing, these models often struggle at low resolutions, limiting their practicality in some real-world scenarios [5].

LiDAR-based gait recognition is an emerging field that utilizes precise 3D representations, point clouds, capturing complex motion patterns and the 3D structure of individual gaits. This approach is less susceptible to variations in lighting, clothing, and background, offering promising avenues for accurate gait recognition in diverse conditions. LidarGait [30] stands as a pioneering work in this field, introducing a multimodal gait recognition dataset named SUSTECH1K. This dataset is designed for evaluating the performance of gait recognition based on various sensors and has demonstrated the potential and practicality of LiDAR-based gait recognition. Additionally, some researchers have applied LiDAR to the task of person re-identification [9], achieving impressive results.

Camera-based and LiDAR-based methods have been proven practical, each with its unique advantages and disadvantages. Therefore, providing a compromise solution by addressing cross-modality gait recognition represents a valuable research direction, which has not been studied currently.

Contrastive Pre-training. Contrastive pre-training is primarily applied in the field of vision-language models, aimed at aligning the feature spaces of visual and linguistic models. In recent years, inspired by the success of self-supervised learning within intra-modal tasks, researchers have begun to explore pre-training objectives for tasks involving multiple modalities, such as vision and language [37]. The pioneering work by CLIP [27] performs cross-modality contrastive pre-training on hundreds of millions of image-text pairs. CLIP [27] can generate a task-agnostic model that achieves surprisingly effective results. Subsequently, ALIGN [14] expands upon CLIP by utilizing a noisy dataset that covers more than a billion image-text pairs, further extending the capabilities of cross-modality pre-training. Contrastive pre-training plays a crucial role in cross-modality tasks. Therefore, we introduce contrastive pre-training into camera-LiDAR cross-modality gait recognition.

RGB-IR Cross-modality Person ReID. Visible-infrared person ReID addresses the challenge of matching persons across different modalities, specifically between RGB and infrared cameras [36]. This task is particularly significant for scenarios with poor illumination, notably at night. With two primary datasets, SYSU-MM01 [36] and RegDB [24], RGB-IR cross-modality person ReID has been extensively studied, most of which focusing on metric learning, feature learning, and adversarial learning approaches. Metric learning [7, 10, 21, 40] and feature learning [25, 39, 41] methods aim to extract and align multimodal features into a common space for effective comparison, employing strategies like angle-based measurement and Euclidean constraints to bridge the modality gap. Notably, advancements in this area include the development of novel loss functions [21]. Concurrently, Generative Adversarial Networks (GANs) [3, 4, 23] have been instrumental in cross-modality ReID, enabling the transformation of images between RGB and IR while maintaining identity information, and fostering feature learning through adversarial and disentanglement strategies. Despite their potential, GAN-based methods face challenges such as high computational demands and the occasional production of low-quality images that may negatively impact ReID performance.

While RGB-IR person ReID mainly relies on appearance for distinction, which is a short-term feature and may raise privacy concerns. In contrast, cross-modality gait recognition focuses on shape and gait of individuals, which are time-invariant features, and offers better protection of privacy.

3 Method

In this work, we propose CL-Gait for cross-modality gait recognition between camera and LiDAR. CL-Gait employs a two-stream network for cross-modality feature embedding. The network utilizes modality-specific modules in the shallow layers and shared modules in the deeper layers, as illustrated in Fig. 2. Besides, CL-Gait adopts a contrastive learning strategy to align the feature spaces of the two modalities, mitigating modality differences, as shown in Fig. 3. The backbone obtained from pre-training is used to initialize the cross-modality embedding network. To facilitate pre-training on large-scale data, we also propose a method for generating cross-modality gait data, as shown in Fig. 4.

3.1 Cross-modality Network of CL-Gait

Two-Stream Backbone. The CL-Gait framework processes two distinct data modalities: 2D silhouette sequences and 3D point cloud sequences, optimizing

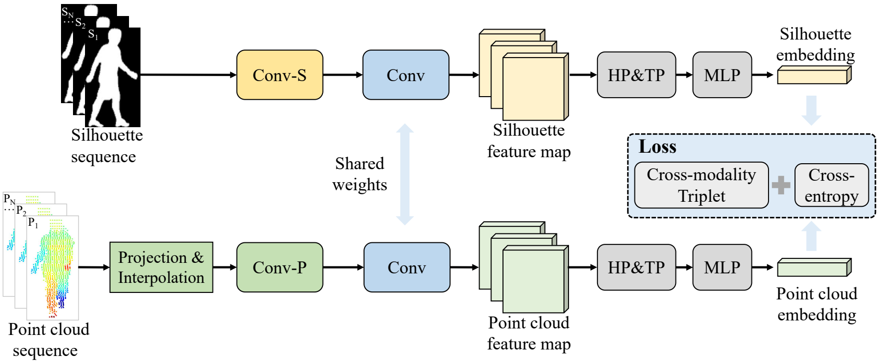


Fig. 2: The cross-modality network of our proposed CL-Gait. It employs a two-stream architecture that encodes sequences from two modalities into a consistent feature space. HP stands for horizontal pooling, and TP represents temporal pooling.

cross-modality feature extracting. Inspired by [30], we first project and interpolate point clouds into depth images. For each modality, we apply modality-specific convolution layers (Conv-S for silhouettes and Conv-P for depth images generated from point clouds) before projecting the features into a shared embedding space, as shown in Fig. 2. Following [36], the two-stream backbone is modified from ResNet-18, with its first layer altered to be the modality-specific convolution layers, and the subsequent layers serving as shared convolution layers. For silhouettes, we define the processing as follows:

$$F_I = L_S(I), \tag{1}$$

where I represents the silhouette and L_S denotes the silhouette-specific convolution layer, which produces the feature map F_I.

Considering the modality differences and the need for consistency in data processing, we first project and interpolate point clouds to generate depths with the same resolution as silhouettes. And the processing is as follows:

$$F_P = L_P(\mathrm{PI}(P)), \tag{2}$$

where P denotes the point cloud, PI represents the projection and interpolation process, and L_P is the depth-specific convolution layer, which produces the feature map F_P for the point cloud.

After extracting modality-specific features F_P and F_S, we aim to unify these features within a consistent feature space. To achieve this, we employ a shared convolutional layer, denoted as L_{sh}, to process both F_P and F_S:

$$F' = L_{\mathrm{sh}}(F_m), \text{ where } F_m = F_P \text{ or } F_S, \tag{3}$$

where F' is the output feature map of L_{sh}.

Temporal and Spatial Pooling. In our network, we follow commonly used strategies in gait recognition for temporal and spatial aggregation, specifically through temporal and horizontal pooling [5]. These techniques are essential for effectively summarizing and interpreting the dynamic and spatial aspects of gait data. Horizontal pooling divides the input feature map horizontally into N_H parts, with each part being aggregated into a single feature vector:

$$\mathrm{HP}: \mathbb{R}^{T\times H\times W\times C} \rightarrow \mathbb{R}^{T\times N_H\times C}, \tag{4}$$

where T denotes the sequence length, H and W represent the height and width of the feature map, respectively, and C indicates the dimension of the features.

Temporal pooling allows for the aggregation of features from variable-length sequences, enabling the model to capture the final sequence-level representation:

$$\mathrm{TP}: \mathbb{R}^{T\times N_H\times C} \rightarrow \mathbb{R}^{N_H\times C}. \tag{5}$$

Therefore, we obtain N_H feature vectors, which are further mapped to the metric space using N_H independent Multi-Layer Perceptrons. These mapped vectors are then concatenated to form the final sequence-level feature embedding.

Training and Inference. The cross-modality network of CL-Gait is trained with cross-modality triplet loss and cross-entropy loss. We have modified the triplet loss to accommodate cross-modality feature learning, where each modality is alternately used as the anchor, while the corresponding positive and negative samples are selected from the other modality, formulated as:

$$\mathcal{L}_{\text{cross-triplet}} = \frac{1}{2}\left(\mathcal{L}_{\text{triplet}}(P, I_{\text{pos}}, I_{\text{neg}}) + \mathcal{L}_{\text{triplet}}(I, P_{\text{pos}}, P_{\text{neg}})\right). \tag{6}$$

The training loss is a combination of both losses with a hyperparameter γ:

$$\mathcal{L} = \mathcal{L}_{\text{cross-triplet}} + \gamma\mathcal{L}_{\text{ce}}, \tag{7}$$

where $\mathcal{L}_{\text{ce}}$ is the cross-entropy loss. During inference, the similarity between samples from different modalities is measured using the Euclidean distance.

3.2 Contrastive Silhouette-point Pre-training

Based on our observations in gait recognition tasks, the significant modality difference between 3D point clouds and 2D images could be a critical factor affecting model performance. To be specific, point clouds focus more on the 3D positioning of body parts, whereas images concentrate on the contour information of the individual. It's crucial to establish connections between the distinct information focused on by each modality. Inspired by CLIP [27], we propose a contrastive silhouette-point pre-training (CSPP) strategy to align the feature spaces of both modalities in the convolution-based encoders, as shown in Fig. 3.

The pre-training process does not require identity labels from the samples for supervision. Trained on paired single-view data from aligned camera and LiDAR, the pre-training could make the model focus on learning a robust representation that bridges the gap between the modalities without direct identity-based guidance, and enhance the performance of the cross-modality network.

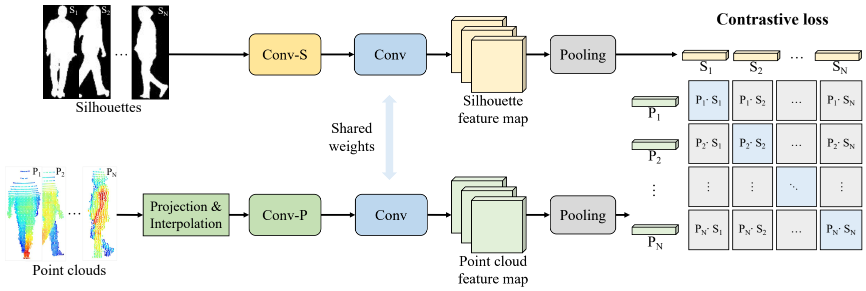


Fig. 3: The contrastive silhouette-point pre-training (CSPP) approach of CL-Gait. Paired silhouettes and point clouds are taken as inputs, and the backbone of the cross-modality network is pre-trained with contrastive learning loss to align the feature spaces of the two modalities. The pre-training process does not require identity labels of the samples for supervision.

Pre-training Process. For pre-training, the input shifts from sequence data to numerous pairs of point clouds and silhouettes. These pairs are defined as data collected from the same individual at the same moment, encompassing both modalities. It ensures that the network learns to reconcile the inherent differences between the 3D and 2D representations, focusing on the alignment of features that accurately reflect the same gait patterns across modalities.

The paired data is processed by the two-stream backbone and spatial pooling operations to obtain paired feature embeddings, i.e., $\mathcal{S} = \{S_i | i = 1, 2..., N\}$ for silhouettes and $\mathcal{P} = \{P_i | i = 1, 2..., N\}$ for point clouds. They are then utilized to compute the loss for contrastive representation learning.

Contrastive Learning loss. The cornerstone of our pre-training is the contrastive learning loss [27]. Given the paired feature embeddings $\mathcal{S} = \{S_i | i = 1, 2..., N\}$ and $\mathcal{P} = \{P_i | i = 1, 2..., N\}$, there are $N \times N$ possible (silhouettes, point clouds) pairings. CL-Gait is trained to learn a multi-modality embedding space by maxmizing the cosine similarity of the feature embeddings of the N real pairs while minimizing the cosine similarity of the embeddings of the $N^2 - N$ incorrect pairings. The cosine similarity matrix is calculated as follows:

$$\mathcal{M} = N_{\mathrm{L2}}(\mathcal{S}) \times N_{\mathrm{L2}}(\mathcal{P})^T, \tag{8}$$

where $\mathcal{S} \in \mathbb{R}^{N \times D}$, $\mathcal{P} \in \mathbb{R}^{N \times D}$, $\mathcal{M} \in \mathbb{R}^{N \times N}$, $N_{\mathrm{L2}}(\cdot)$ denotes L2 normalization, and $(\cdot)^T$ denotes the transpose of the matrix. Then, a symmetric cross entropy loss over these similarity scores is optimized:

$$\mathcal{L}_{\mathrm{con}} = \frac{1}{2} \left(\mathcal{L}_{\mathrm{ce}}(\mathcal{M}, G) + \mathcal{L}_{\mathrm{ce}}(\mathcal{M}^T, G) \right), \tag{9}$$

where $G = \{1, 2..., N\}$. G represents the labels used for calculating the cross-entropy loss, indicating that pairs corresponding to the similarities on the diagonal of matrices $\mathcal{M}$ and $\mathcal{M}^T$ are considered positive samples, while all others are treated as negative samples. This approach ensures that the network learns to distinguish between matching and non-matching modality pairs effectively.

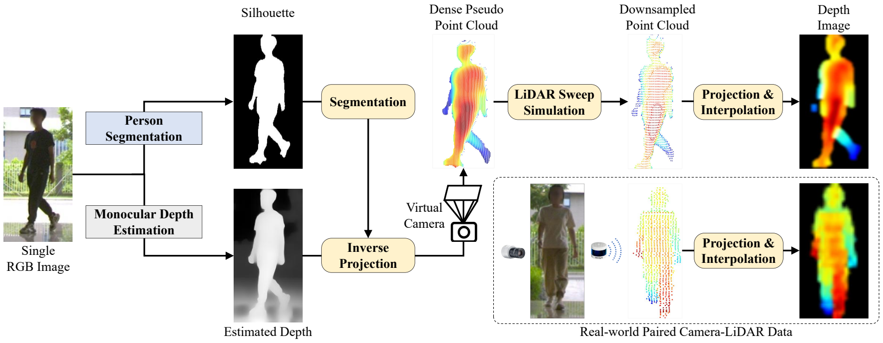


Fig. 4: Illustration of paired gait data generation from single RGB images for contrastive pre-training. The quality of the synthesized data is comparable to the real-world data, making it possible to synthesize large-scale pre-training data.

In the process described above, each frame of point cloud or silhouette, after passing through the pre-trained network, yields a feature embedding of size $(1, C)$, indicating global pooling is applied to the output feature map of the backbone. For practical application and to align each local feature embedding, horizontal pooling is used to obtain features of size (N, C) for each frame. During the computation of the contrastive learning loss, pairs of features from the same individual, moment, and part across two modalities are considered positive samples, with all others being negative. Through our practice, we have found that adopting a strategy of aligning local features during the pre-training phase is more suitable for gait recognition tasks, leading to performance improvements.

3.3 Generation of Pre-training Gait Data

Pre-training on real and large-scale data is challenging due to the high cost of acquiring paired RGB and point cloud data. To address this issue, we propose a method for generating pseudo data based on monocular depth estimation. As illustrated in Fig. 4, we use Depth Anything [38] to estimate dense depth $D \in \mathbb{R}^{H \times W}$ from large-scale single RGB images. Then, utilizing a virtual camera with intrinsic K, the estimated depth is mapped to pseudo point clouds:

$$[P', 1]^T = K^{-1} D [u, v, 1]^T, \tag{10}$$

where $P' \in \mathbb{R}^{H \times W \times 3}$ is the point coordinates, K^{-1} indicates the inverse of intrinsic K, and u, v are the image pixel coordinates. The pseudo points P', after being downsampled by voxel grid into P'_{down}, can be used for contrastive pre-training between camera silhouettes and LiDAR point clouds. To be specific, the downsampled points P'_{down} are projected back into the image pixel coordinate to obtain the depth images for contrastive pre-training. Compared to point clouds, RGB images are less costly to collect and easier to acquire, with many public datasets of pedestrian image already available. Our proposed method makes it feasible to synthesize large-scale data for contrastive pre-training.

4 Experiments

4.1 Datasets

The SUSTech1K Dataset [30] is the first large-scale LiDAR-based gait recognition dataset collected by a LiDAR sensor and an RGB camera. The dataset contains 25,239 sequences from 1,050 subjects and covers many variations, including visibility, views, occlusions, clothing, carrying, and scenes. It is timestamped frames for each modality of frames. The first single-modality 3D gait recognition framework is trained on SUSTech1K using point clouds. In our study, we explore new uses for the SUSTech1K dataset. Our approach involves both images and point clouds, which marks the first attempt of cross-modality gait recognition with this dataset. We also introduce a novel pre-training strategy based on paired images and point clouds from the dataset. This strategy utilizes contrastive learning for cross-modality matching between images and point clouds and improves the performance of our proposed cross-modality gait method.

The HITSZ-VCM Dataset and LIP Dataset are large-scale datasets, which contains a large amount of person videos/images captured by cameras. In detail, the HITSZ-VCM dataset [20] contains 927 valid identities with image sequences. The LIP dataset [8] contains about 50,000 identities but with only one image for each. Totally, there are 251,452 RGB images in the HITSZ-VCM dataset and 50,000 in the LIP dataset. We implement the pre-training strategy on these datasets containing only RGB images, to explore the impact of pre-training with out-of-domain data on this task. Given that there is no point cloud data paired with these images, we use monocular depth estimation to generate pseudo point clouds, as shown in Fig. 4. This greatly expands the amount of data available for pre-training at scale. This also improves the accuracy of cross-modality recognition between 2D and 3D spaces, which marks the effectiveness of our proposed contrastive pre-training strategy.

4.2 Experimental Setup

Implementation Details. Following [30], all the camera-based silhouettes and LiDAR-based depth images are aligned and then resized into the resolution of 6464. The total number of iterations is set to 120,000 for pre-training and 60,000 for fine-tuning. The Adam optimizer is used to prevent the issue of gradient vanishing because of low-quality silhouettes. The triplet and cross-entropy loss weights are set to the same. All comparison methods are trained using the same training strategy as LidarGait [30]. The OpenGait codebase [5] is utilized to conduct all experiments.

Evaluation Protocol. All experiments are conducted on SUSTech1K. Following LidarGait [30], the dataset is divided into three splits: a training set with 250 identities and 6,011 sequences, a validation set with 6,025 sequences from 250 unseen identities, and a test set with the remaining 550 identities and 13,203 sequences. The SUSTech1K dataset provides gait sequences from multiple viewpoints. Thus, we also adopt the cross-view evaluation protocol [31, 42] used in

CASIA-B and OUMVLP. During the test, the sequences in normal conditions are grouped into gallery sets, and the sequences in variant conditions are taken as probe sets. To accurately assess the performance of camera-LiDAR cross-modality gait recognition, we evaluate the results using point cloud data as the probe and silhouette data as the gallery, i.e., LiDAR to Camera (L to C). We also evaluate the results using silhouette data as the probe and point cloud data as the gallery, i.e., Camera to LiDAR (C to L).

4.3 Comparative Methods

We evaluate several commonly used structures for cross-modality modeling as baselines, which basically follows the setting of [36]. The baseline models includes one-steam structure, asymmetric FC layer structure, image-point structure, and the proposed two-stream structure. We apply residual block in ResNet-18 [11] as the base convolution block for all the four structures. For them, the loss function are weighted cross-entropy loss and triplet loss [2], which is commonly used and relatively stable. And all of the hyper parameters are kept the same. As for the input of one-steam structure, asymmetric FC layer structure and two-stream structure, the point clouds are converted to three-channel colored depth images and adjusted to the resolution of 64 × 64 with zero padding. We introduce the four structures as following:

One-stream Structure is the most commonly used in vision tasks, where all parameters are shared in the whole network. Therefore, we organize silhouettes and depth images obtained from point clouds into the same dimensions before feeding them into the network.

Asymmetric FC Layer Structure is used in multi-domain tasks, for example, IDR [12] for VIS-NIR face recognition. This structure shares nearly all parameters except for the last FC layer. It assumes that the feature extraction for different modalities can be same and adaptation is achieved in the backbone.

Image-point Structure directly uses point clouds as input, rather than projected depth images. It employs two types of encoders: an image encoder for the silhouette input and a point cloud encoder for the point cloud input. Under this structure, we implement three commonly used point cloud encoders: GCN [9], PointNet [26], and Point-Transformer [43].

Two-stream Structure is commonly used in cross-modality matching tasks. It utilizes modality-specific modules in the shallow layers of the network and uniform modules with shared parameters in the deeper layers. Based on the two-stream structure, our proposed CL-Gait is equipped with a horizontal pooling module (HP) and a contrastive silhouette-point pre-training strategy (CSPP). To show the effectiveness of the HP module and CSPP strategy of CL-Gait, we implement three additional networks: 1) a two-stream backbone without CSPP, using global pooling instead of HP; 2) a two-stream backbone without CSPP, using HP; 3) a two-stream backbone with CSPP, using global pooling.

Table 1: Evaluation with different structures on SUSTech1K valid + test set. We use ResNet-18 to extract image features by default. "L to C" indicates the results with point clouds as probe and silhouettes as gallery, and "C to L" indicates the reverse.

Structure	**Model**	**L to C**			**C to L**		
		Rank-1	**Rank-3**	**Rank-5**	**Rank-1**	**Rank-3**	**Rank-5**
One-stream	ResNet	2.60	5.61	7.99	14.94	25.92	31.87
Asymmetric FC	ResNet	3.08	7.23	10.52	16.43	29.13	35.90
Image-point	GCN [9]	12.20	25.01	33.23	14.30	28.38	36.62
	PointNet [26]	7.11	16.61	23.35	8.40	19.25	26.90
	Point-Transformer [43]	13.25	26.72	35.29	15.42	30.04	38.70
Two-stream	ResNet	30.39	48.77	57.31	35.26	54.42	62.84
	ResNet+CSPP	32.18	50.85	59.22	36.54	55.75	64.11
	ResNet+HP	45.41	62.89	69.71	42.30	59.67	67.12
	CLGait (ours)*	**53.29**	**69.54**	**75.59**	**55.12**	**71.23**	**77.31**

* i.e., ResNet+CSPP+HP.

4.4 Comparative Results

Following the cross-view and cross-modality evaluation protocol, we report the comparative results of the methods above in the Tab. 1. From the results, we can obtain the following observations: 1) CL-Gait demonstrates its superiority to all existing structures and methods, primarily due to the feature extraction capabilities of the two-stream structure. Additionally, the use of pre-training effectively mitigates modality discrepancy, further improving its performance. 2) Experiments under the two-stream structure can be considered as an ablation study, highlighting the important roles of CSPP and HP. Adding HP to the ResNet base improves the average rank-1 by approximately 11%, and further incorporating CSPP increases the average rank-1 by about another 10%. 3) The performance of the one-stream structure is poor, primarily because the one-stream structure uses an identical network to process both silhouettes and point clouds, failing to address the modality discrepancy between them. 4) Although the Asymmetric FC structure employs different FC layers for the two modalities, its performance remains poor. This indicates that modality-specific adaptations are best applied at the shallower levels of the network. Furthermore, during the training process, the asymmetric FC layers tend to cause the network to overfit quickly. 5) All methods utilizing the two-stream structure outperform those based on the image-point structure. This likely results from the depth images, obtained through projection, having data consistency and smaller modality discrepancy with silhouettes compared to point cloud inputs.

4.5 Discussion

3D Geometry Information. To evaluate the effectiveness of depth information for cross-modality gait recognition, we compare another type of data as input for the second modality, e.g., LiDAR silhouettes. LiDAR silhouettes are obtained by range-view projection of point cloud sets and contains no 3D geometry information. The results are shown in Fig. 5. When 3D geometry information is not

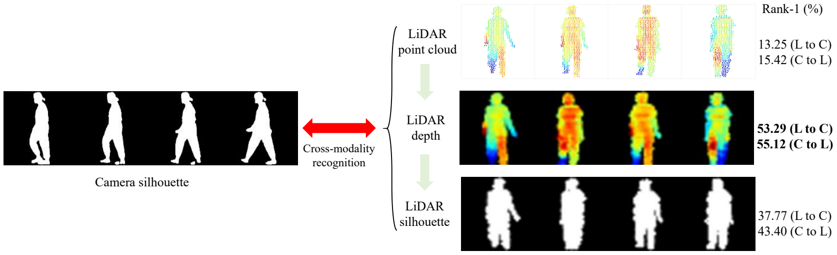


Fig. 5: Comparison of different input forms of LiDAR data. The results show that the projected and interpolated depth from point cloud works best for the cross-modality matching. This indicates that the 3D geometry information are essential. For each input form, the best-performing model is utilized.

included, the performance of LiDAR silhouettes is much lower. This is because the camera has a much higher resolution than LiDAR, so the silhouettes from the camera can have more details. Integrating depth information can improve average rank-1 accuracy from 40.59% to 54.21%, validating the necessity and effectiveness of 3D information for cross-modality gait recognition. One important reason may be that the 3D information contained in the point cloud can compensate for the difference in viewing angles to some extent. Furthermore, we include the performance of direct point cloud input in Fig. 5 to compare with depth images and illustrate the input form of point clouds.

Contrastive Pre-training Datasets. To make up for the absence of paired camera-LiDAR data, we further propose a large-scale data generation strategy. Camera silhouette and LiDAR point data are generated from real-world RGB images for contrastive pre-training on such paired pseudo data. To valid its performance, we construct training sets from two different datasets, including LIP and HITSZ-VCM. These two datasets are commonly used for camera-based person segmentation and re-identification and contain single still RGB images. To ensure fair comparisons, we also conduct experiments on SUSTech1K training set, which contains real paired point clouds and camera silhouettes.

As shown in Tab. 2, when the amount of data is similar using data generated from the LIP dataset, the performance of pre-training on generated pseudo data is comparable to that of real data. When pre-trained on a larger scale of synthetic pseudo data generated from the HITSZ-VSM dataset, the accuracy from LiDAR to camera improves and exceeds that on the real data, proving the effectiveness of our data generation strategy for contrastive pre-training. However, the accuracy from camera to LiDAR decreases a little, probably due to the difference between generated point clouds from estimated depth and real points from LiDAR.

Temporal Modules. In order to aggregate the different frames in the sequence, temporal modules are used to fuse the features of multiple frames. We evaluate the effectiveness of different temporal modules in combination with our proposed CL-Gait, as shown in Tab. 3. We can observe that the simple temporal pooling structure demonstrates superiority over the other learning-based structure. This

Table 2: Performance with different pre-training datasets. Using estimated depth for pre-training obtains close results compared with using real paired point-image dataset.

Dataset	Data Amount	L to C			C to L		
		Rank-1	Rank-3	Rank-5	Rank-1	Rank-3	Rank-5
SUSTech1K	120K	51.60	68.22	74.43	**56.12**	**72.15**	**77.95**
LIP	50K	51.00	68.09	73.99	55.07	71.59	77.59
HITSZ-VCM	770K	**53.29**	**69.54**	**75.59**	55.12	71.23	77.31

Table 3: Performance of CL-Gait with different temporal modules.

Feature	L to C			C to L		
	Rank-1	Rank-3	Rank-5	Rank-1	Rank-3	Rank-5
Temporal Pooling [5]	**53.29**	**69.54**	**75.59**	**55.12**	**71.23**	**77.31**
LSTM [13]	37.40	55.01	60.89	39.47	56.41	62.72
Bi-LSTM [28]	39.63	55.55	61.83	39.88	57.82	63.71
Transformer [33]	44.80	61.31	67.36	47.16	63.83	69.11

result indicates that temporal pooling can better fuse multi-frame features to match domain information more effectively. One reason may be that learning-based approaches such as LSTM and Transformer are more likely to overfit.

5 Conclusion and Future work

This paper presents the first research on cross-modality gait recognition with point clouds and silhouettes from RGB images. Firstly, We propose a cross-modality gait recognition framework, named CL-Gait, that utilizes a two-stream network for feature embedding of different modalities, i.e., camera and LiDAR. Moreover, to mitigate modality discrepancy, we propose a contrastive pre-training strategy along with a large-scale data generation method. It can generate large number of data pairs (silhouette and point cloud) based on monocular depth estimation and pre-train the model in a contrastive manner. Extensive experiments demonstrate that CL-Gait achieves commendable performance in cross-modality evaluation modes, proving the importance of maintaining modality consistency and the effectiveness of contrastive pre-training. This also highlights the potential of cross-modality gait recognition.

CL-Gait has shown remarkable results, but we believe there remains significant room for improvement. The modality and resolution differences between point clouds and silhouettes, along with the inherent motion blur in point clouds, may be key factors impacting model performance. Therefore, densifying point clouds or developing point cloud encoders specifically for cross-modality gait recognition could enhance model performance.

Acknowledgement

This work is supported by the National Key Research and Development Program of China under Grant (2018AAA0102803).

References

1. Bobick, A.F., Davis, J.W.: The recognition of human movement using temporal templates. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(3), 257–267 (2001)
2. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8126–8133 (2019)
3. Choi, S., Lee, S., Kim, Y., Kim, T., Kim, C.: Hi-cmd: Hierarchical cross-modality disentanglement for visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10257–10266 (2020)
4. Dai, P., Ji, R., Wang, H., Wu, Q., Huang, Y.: Cross-modality person re-identification with generative adversarial training. In: IJCAI. vol. 1, p. 6 (2018)
5. Fan, C., Liang, J., Shen, C., Hou, S., Huang, Y., Yu, S.: Opengait: Revisiting gait recognition towards better practicality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9707–9716 (2023)
6. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14225–14233 (2020)
7. Feng, Z., Lai, J., Xie, X.: Learning modality-specific representations for visible-infrared person re-identification. IEEE Transactions on Image Processing **29**, 579–590 (2019)
8. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 932–940 (2017)
9. Guo, W., Pan, Z., Liang, Y., Xi, Z., Zhong, Z.C., Feng, J., el al: Lidar-based person re-identification. arXiv preprint arXiv:2312.03033 (2023)
10. Hao, Y., Wang, N., Li, J., Gao, X.: Hsme: Hypersphere manifold embedding for visible thermal person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 8385–8392 (2019)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
12. He, R., Wu, X., Sun, Z., Tan, T.: Learning invariant deep representation for nir-vis face recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
13. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8), 1735–1780 (1997)
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., el al: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021)
15. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: Proceedings of the Asian Conference on Computer Vision (2020)
16. Liang, J., Fan, C., Hou, S., Shen, C., Huang, Y., Yu, S.: Gaitedge: Beyond plain end-to-end gait recognition for better practicality. In: European Conference on Computer Vision. pp. 375–390. Springer (2022)

17. Liao, R., Yu, S., An, W., Huang, Y.: A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognition **98**, 107069 (2020)
18. Lin, B., Liu, Y., Zhang, S.: Gaitmask: Mask-based model for gait recognition. In: BMVC. pp. 1–12 (2021)
19. Lin, B., Zhang, S., Wang, M., Li, L., Yu, X.: Gaitgl: Learning discriminative global-local feature representations for gait recognition. arXiv preprint arXiv:2208.01380 (2022)
20. Lin, X., Li, J., Ma, Z., Li, H., Li, S., Xu, K., Lu, G., Zhang, D.: Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20973–20982 (2022)
21. Ling, Y., Zhong, Z., Luo, Z., Rota, P., Li, S., Sebe, N.: Class-aware modality mix and center-guided metric learning for visible-thermal person re-identification. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 889–897 (2020)
22. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. In: Seminal Graphics Papers: Pushing the Boundaries, Volume 2, pp. 851–866 (2023)
23. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., Yu, N.: Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13379–13389 (2020)
24. Nguyen, D.T., Hong, H.G., Kim, K.W., Park, K.R.: Person recognition system based on a combination of body images from visible light and thermal cameras. Sensors **17**(3), 605 (2017)
25. Park, H., Lee, S., Lee, J., Ham, B.: Learning by aligning: Visible-infrared person re-identification using cross-modal correspondences. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12046–12055 (2021)
26. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 652–660 (2017)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
28. Schuster, M., Paliwal, K.K.: Bidirectional recurrent neural networks. IEEE Transactions on Signal Processing **45**(11), 2673–2681 (1997)
29. Sepas-Moghaddam, A., Etemad, A.: Deep gait recognition: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 264–284 (2022)
30. Shen, C., Fan, C., Wu, W., Wang, R., Huang, G.Q., Yu, S.: Lidargait: Benchmarking 3D gait recognition with point clouds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1054–1063 (2023)
31. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Transactions on Computer Vision and Applications **10**, 1–14 (2018)
32. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: Graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP). pp. 2314–2318. IEEE (2021)

33. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in Neural Information Processing Systems **30** (2017)
34. Wang, C., Zhang, J., Pu, J., Yuan, X., Wang, L.: Chrono-gait image: A novel temporal template for gait recognition. In: Computer Vision–ECCV 2010: 11th European Conference on Computer Vision, Heraklion, Crete, Greece, September 5-11, 2010, Proceedings, Part I 11. pp. 257–270. Springer (2010)
35. Wang, Y., Chao, W.L., Garg, D., Hariharan, B., Campbell, M., Weinberger, K.Q.: Pseudo-lidar from visual depth estimation: Bridging the gap in 3D object detection for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8445–8453 (2019)
36. Wu, A., Zheng, W.S., Yu, H.X., Gong, S., Lai, J.: Rgb-infrared cross-modality person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5380–5389 (2017)
37. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., Zeng, B., Chilimbi, T., Huang, J.: Vision-language pre-training with triple contrastive learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15671–15680 (2022)
38. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. arXiv preprint arXiv:2401.10891 (2024)
39. Ye, M., Lan, X., Leng, Q.: Modality-aware collaborative learning for visible thermal person re-identification. In: Proceedings of the 27th ACM International Conference on Multimedia. pp. 347–355 (2019)
40. Ye, M., Lan, X., Leng, Q., Shen, J.: Cross-modality person re-identification via modality-aware collaborative ensemble learning. IEEE Transactions on Image Processing **29**, 9387–9399 (2020)
41. Ye, M., Shen, J., J. Crandall, D., Shao, L., Luo, J.: Dynamic dual-attentive aggregation learning for visible-infrared person re-identification. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16. pp. 229–247. Springer (2020)
42. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (ICPR'06). vol. 4, pp. 441–444. IEEE (2006)
43. Zhao, H., Jiang, L., Jia, J., Torr, P.H., Koltun, V.: Point transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16259–16268 (2021)