Robust Gait Recognition Based on Deep CNNs With Camera and Radar Sensor Fusion

Yu Shi, *Graduate Student Member, IEEE*, Lan Du, *Senior Member, IEEE*, Xiaoyang Chen, Xun Liao, Zengyu Yu, Zenghui Li, Chunxin Wang, and Shikun Xue

***Abstract*—In recent years, gait recognition has emerged as an important and promising solution for human identification. Generally, gait recognition is based on a single type of sensor, such as a camera or a radar. However, data of a single modality may only capture inadequate gait features of a person, such as camera data lacking the intuitive micro-motion pattern information and radar data lacking the information about gait appearance, making gait-based human identification system vulnerable to complex covariate conditions, e.g., cross-view and cross-walking-condition. To build a robust and reliable gait-based human identification system, in this study, we propose a multisensor gait recognition framework with deep convolutional neural networks (CNNs) by fusing camera gait energy images (GEIs) and radar time-Doppler spectrograms. To learn the fine-grained gait appearance features, we propose a body-part spatial attention (BPSA) module to obtain more discriminative body part representations of GEIs. To learn the gait micro-motion pattern, we propose a long-short temporal relation modeling (LSTRM) module to obtain the local and global micro-motion representation of time-Doppler spectrograms. Finally, we fuse the discriminative body part representation and the micro-motion pattern at the multiscale feature space to obtain richer and more robust gait features for human identification. We provide an extensive empirical evaluation in terms of various complex covariate conditions, namely, cross-view and cross-walking-condition. Experiments on 121 subjects with eight views and three walking conditions of camera and radar data show our proposed method is more robust and accurate.**

***Index Terms*—Body-part attention, camera gait energy images (GEIs), cross-view, cross-walking-condition, deep convolutional neural networks (CNNs), gait recognition system, multisensor fusion, radar time-Doppler spectrograms, temporal relation modeling.**

I. Introduction

GAIT recognition is an appealing biological identification technology, which aims to identify humans based on their unique walking ways [1]. Because gait is difficult to disguise and deceive, it has been widely applied in smart space, such as crime prevention, social security, access control for automatic doors and IoT devices, and so on [2].

Manuscript received 10 October 2022; accepted 23 January 2023. Date of publication 16 February 2023; date of current version 7 June 2023. This work was supported in part by the National Science Foundation of China under Grant U21B2039, and in part by the 111 Project. *(Corresponding author: Lan Du.)*

The authors are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: shiyu1213@126.com; dulan@mail.xidian.edu.cn; xiaoyangchen@stu.xidian.edu.cn; xliao@stu.xidian.edu.cn; bithogan@163.com; zhli_1@stu.xidian.edu.cn; wangchunxin1126@163.com; skxue@stu.xidian.edu.cn).

Digital Object Identifier 10.1109/JIOT.2023.3242417

Potential sources for gait biometrics can be seen as coming from two aspects: 1) appearance and 2) motion [3]. Unlike coarse-grained human activity recognition, gait recognition depends on the subtle discriminative gait information (appearance and motion), i.e., micro-pattern, which is often more challenging than those dealt with for activity recognition [2]. There are a number of sensors that can capture gait information, such as camera, radar, accelerometer, force, and pressure sensors [4], [5], [6], [7], [8]. Among them, the camera can easily capture gait appearance information but cannot effectively capture motion cues, especially, micro-motion patterns [4]. Meanwhile, the radar, accelerometer, etc., are good at portraying the gait motion pattern but cannot effectively capture appearance cues [4], [9]. It is worth noting that the radar can further capture micro-motion pattern information, such as swings of human arms, legs, feet, and other parts [9], [10]. Multisensor can capture gait patterns from different aspects and provide complementary information on gait. However, most current gait recognition systems are based on a single type of sensors, and data of a single modality may only capture inadequate gait features of a person, which limits the accuracy and robustness of gait-based human identification systems, especially, under complex covariate conditions. The challenging complex covariate conditions that affect the performance of gait recognition mainly include the following two aspects [2], [11], [12], [13]: 1) variations in the sensor's viewpoint or subject's viewpoint, i.e., cross-view condition and 2) variations in the walking conditions of the subjects, such as carrying a bag or wearing items of clothing such as a coat, i.e., cross-walking-condition. In this article, we mainly focus on gait recognition systems based on camera and radar, since camera and radar are two popular nonwearable sensors that have the advantage of being able to capture gait information from long distances without human cooperation compared with other wearable sensors. This advantage is also not available with other traditional biometric features, such as face, fingerprint, and iris, making gait recognition more promising for a wider range of applications. We aim to fuse the camera and radar gait information to improve the gait recognition accuracy and robustness under complex covariate conditions.

Camera-based methods are the most extensively and deeply studied in the gait recognition community. Due to the powerful ability to automatically learn discriminative representations, deep convolutional neural networks (CNNs) methods currently dominate the field and facilitate real-world applications in this area [2]. Since the camera can

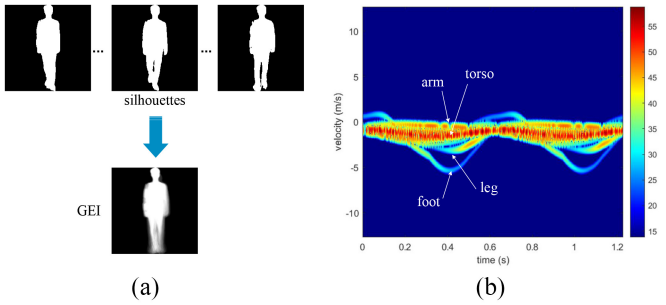


Fig. 1. Format illustrations of camera data and radar data. Camera data can reflect gait appearances but lack the intuitive micro-motion pattern information, while radar time-Doppler spectrogram contains effective information about the human micro-motion of the body part. (a) Format illustrations of camera data. (b) Illustration of radar time-Doppler spectrogram.

easily capture gait appearance information, the camera-based approach is mainly used to extract discriminative representations of gait appearance or gait appearance change over time. There are two main ways to recognize the gait, as shown in Fig. 1(a), one is regarding gait as a silhouette sequence that is extracted by human detection and segmentation from the raw video sequence [14], [15], [16], [17], and the other is regarding gait as a gait template by compressing the gait sequence into a single image, for instance, gait energy images (GEIs) [18], [19], [20]. The advantage of the gait template is simple and easy to implement, but it tends to lose time information. Identifying the gait from the gait silhouette sequences can better capture gait information. However, deep neural networks like 3D-CNNs for extracting sequential information are harder to train and deploy to real-world systems than those using a single template since these models generally have larger parameters and computational cost [14], [19]. Despite the tremendous progress of the camera-based deep gait recognition methods, the inability of the cameras to effectively capture motion cues, especially, micro-motion pattern information, makes the performance of these methods still limited, especially, under complex covariate conditions, e.g., cross-view and cross-walking-condition.

Since the micro-Doppler effect can reflect the micro-motion of humans, gait recognition methods based on time-Doppler spectrograms (a form of radar data) have attracted a lot of attention from researchers in recent years. The micro-Doppler effect refers to the physical phenomenon of Doppler frequency modulation of the radar echo signal by the rotation, vibration, oscillation, and other micro-motion of targets [21], [22], [23], [24]. As shown in Fig. 1(b), the human micro-motion characteristics contain effective information about the human motion state and body posture, such as swings of human arms, legs, feet, and other parts. Since time-Doppler spectrograms are well suitable for the gait recognition task [25], time-Doppler-based gait recognition methods have been investigated recently. In these researches, deep learning methods, especially, CNNs, have achieved promising performance [26]. Nevertheless, the performance of current radar gait recognition methods is still limited due to the two following reasons: 1) compared to appearance information, micro-Doppler information is difficult to be interpreted and 2) radar data lack gait appearance information. Besides that, existing radar-based methods are only limited to ideal conditions and do not explore gait recognition under cross-view and cross-walking-condition.

Camera and radar can capture highly complementary information for gait recognition, which can alleviate the different kinds of errors and the sensitivity to complex covariate conditions introduced by a single type of sensor. Motivated by this and sophisticated developments in deep learning methods, in this article, we propose a robust gait recognition approach with deep CNNs for complex covariate conditions by fusing the camera and radar sensors. We use an RGB camera to collect GEIs and use a millimeter-wave radar to collect time-Doppler spectrograms. The GEIs portray a compact representation of the spatial body appearance over the gait cycle, while the time-Doppler spectrograms provide the micro-motion pattern of gait. To learn the fine-grained gait appearance features and the gait micro-motion pattern, we design a two-stream CNN and propose two new modules to extract gait features from the GEIs data and the time-Doppler spectrograms data, respectively. Specifically, we propose a body-part spatial attention (BPSA) module to obtain more discriminative body part representations of GEIs. Meanwhile, we propose a long-short temporal relation modeling (LSTRM) module to obtain the local and global micro-motion representation of time-Doppler spectrograms. Finally, we fuse the discriminative body part representation and the micro-motion pattern at the multiscale feature space to obtain richer and more robust gait features for human identification. In the experiments, we collect camera and radar data from 121 subjects with eight views and three walking conditions. We further provide an extensive empirical evaluation in terms of various complex covariate conditions, namely, cross-view and cross-walking-condition. Comparison results with other approaches demonstrate that the gait recognition accuracy and robustness can be improved by fusing camera and radar data. More specifically, the main contributions of the proposed method are as follows.

1) We propose a robust gait recognition approach for complex covariate conditions by fusing the camera and radar sensors. Considering the problem that a single type of sensor may only capture inadequate gait features of a person, we fuse the discriminative appearance representation and the micro-motion pattern of camera and radar data to obtain richer and more robust gait feature representations. The multisensor fusion leads to more robust and accurate gait-recognition performance.
2) We propose two new modules to, respectively, extract gait features from the GEIs data and the time-Doppler spectrograms data to learn the fine-grained gait appearance features and the gait micro-motion pattern.
3) We collect 121 subjects with eight views and three walking conditions of camera and radar data and further provide an extensive empirical evaluation in terms of various complex covariate conditions, namely, cross-view and cross-walking-condition. Comparison results with other methods show great potential for practical use of our method in terms of high accuracy and robustness.

In the remainder of this article, Section II introduces the related work on gait recognition. Section III will demonstrate the details of the proposed method, including system overview, data processing, and robust gait recognition with multisensor fusion. Section IV presents the experiments and analysis. Finally, the conclusion is drawn in Section V.

II. Related Work

In this section, we will discuss different gait recognition methods based on single-modality and multimodality.

A. Single-Modality-Based Gait Recognition Methods

There are a number of sensors that can capture gait information, such as camera, radar, accelerometer, force, and pressure sensors. For single-modality, we mainly focus on the camera- and radar-based gait recognition methods.

1) Camera-Based Gait Recognition Methods: Due to the powerful ability to automatically learn discriminative representations, deep CNNs methods currently dominate the field and facilitate practical applications in this area. In the real-world, gait recognition is often regarded as a retrieval task, that is to say, given the samples to be tested, i.e., probe samples, gait recognition system needs to find the best match from the gallery. The pipeline of typical camera-based deep gait recognition methods generally can be described as follows [2], [19]: 1) these methods first obtain the human silhouettes of each frame or gait template such as GEIs; 2) they then extract the gait representation from the silhouette sequence or gait template by deep networks; 3) they measure the similarity or distance between representation pairs of probe and gallery by some metric learning methods; and 4) they finally assign the human identity by nearest neighbor classifier or other classifiers. These methods generally can be grouped into template-based and sequence-based categories. Sequence-based methods extracted gait features from the sequence of binary silhouettes. Lin et al. [17] designed multiple-temporal scale 3-D CNNs to extract richer gait features. Zhang et al. [27] introduced LSTM module to learn the importance weights for each frame and extract more discriminative features. GaitSet [14] utilized gait as a deep set and extracted gait features from a gait frame set to identify gaits. GaitPart [15] is proposed to learn the visual appearances and movement patterns of different body parts. Based on the GateSet, GLN [28] leverages the inherent feature pyramid to enhance the gait representations. GaitGL [29] proposes a new feature extraction and fusion framework to achieve discriminative feature representation for gait recognition. Specifically, GaitGL develops a global and local feature extractor using global visual information and local area details. CSTL [16] introduces relation modeling among multiscale features to enhance the feature representation and proposes a salient spatial feature learning module to tackle the misalignment problem caused by the temporal operation. Wu et al. [19] extracted gait features based on gait templates, e.g., GEIs [18], by 2-D CNNs, which greatly reduced the computation cost. Due to the advantage of the GEIs being simple and easy to implement, our approach takes the GEIs as input of the camera branch.

Despite the tremendous progress of the camera-based gait recognition methods, these methods still face great challenges under complex covariate conditions, especially, cross-view and cross-walking-conditions. Muramatsu et al. [30] used a view transformation model to map the gait features from different view angles to a common one. Vandersmissen et al. [9] aims to project the original gait features to subspace to extract view-invariant features. Verlekar et al. [31] attempted to find the representations of body parts not related to walking conditions. Other methods aim to mitigate the effects of clothing changes on gait recognition by constructing different variants of GEIs, such as gait period dependent images [32], edge-masked active energy images [33], and 5/3 gait images [34].

However, the inability of the cameras to effectively capture motion cues, especially, micro-motion pattern information, makes the performance of these methods still limited, especially, under complex covariate conditions.

2) Radar-Based Gait Recognition Methods: In recent years, some researchers have investigated radar-based gait recognition by extracting time-Doppler gait features. Traditional methods generally extract handcraft features [35], [36], [37], such as time, frequency, and Doppler bandwidth, to classify human identity. These methods require professional knowledge and have a poor performance. Deep CNNs have reshaped the research landscape in the time-Doppler-based gait recognition area. Cao et al. [38] first proposed a gait recognition method based on Deep CNNs for time-Doppler spectrograms, which achieve about 85.6% performance with ten subjects. Lang et al. [39] designed a multitask model to classify actions and persons simultaneously, which achieve an identification accuracy of 80.92% with 15 subjects. Yang et al. [40] further proposed a CNNs with multilayer fusion to improve gait recognition performance. Ozturk et al. [25] proposed the gaitcube to improve gait recognition accuracy and reduce the training overhead. Gaitcube achieves 98.3% with ten persons. Addabbo et al. [41] used the temporal CNNs to model the short-term temporal motion information, which achieves the better performance. Despite the powerful feature extraction capabilities of deep CNNs, these methods do not take full advantage of the long-term temporal motion information of the time-Doppler spectrograms. Meanwhile, existing researches focus only on the classification task, i.e., the identities of the person in the training and test sets have to be consistent, which limits the application of radar-based gait recognition in the real world. As mentioned in the previous section, for practical applications, the gait recognition is often regarded as a retrieval task. Taking a school or factory access control system as an example, compared with classification-based gait recognition, retrieval task-based gait recognition has the following advantages.

1) The retrieval task does not require the identities of people in the training and test sets to be identical, which means that an offline trained gait recognition system can be deployed directly to a new factory or school.

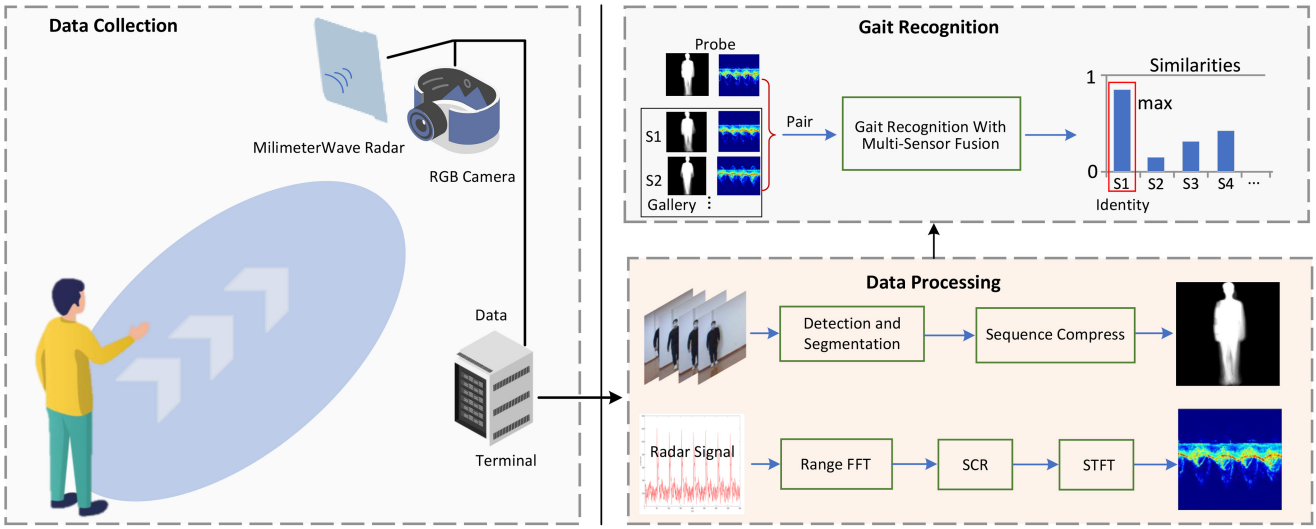


Fig. 2. Overview of our gait-based human identification system with multisensor fusion, which includes data collection, data processing, and gait recognition.

2) When a new person is onboarded, the retrieval task-based gait recognition model only needs to update the gallery while the classification-based gait recognition needs to be retrained. Cheng and Liu [42] explored the retrieval-based gait recognition for radar point clouds data, achieving surprising results. However, point clouds lose the micro-Doppler information of gait for a certain, which limits the performance. Besides that, existing radar-based methods are limited to ideal conditions and do not explore gait recognition under cross-view and walking conditions.

3) Gait Recognition Methods of Other Sensors: There are also some gait recognition methods using other sensors, such as WiFi and accelerometer. WiDIGR [43] proposes a series of signal processing techniques to eliminate the differences among induced WiFi signals caused by walking in different directions and generate a high-quality direction-independent signal spectrogram, achieving promising classification results. Based on WiFi signal, Cao et al. [44] proposed a novel Balloon mechanism to achieve lightweight. CAUTION [45] provides an effective way to train the system for unknown intruder detection and does not require a large number of training WiFi data. Sun et al. [46] proposed the speed-adaptive gait cycle segmentation method and individualized matching threshold generation method to address the problem of the speed of human changes in accelerometer signal. Zhang et al. [5] detected signature-meaningful points on the acceleration curve and extracts gait features for gait recognition. These methods solve the problem of gait recognition from different aspects and improve greatly. However, most of these methods focused on the classification task, which limits the application of these methods in the real world.

B. Multimodality-Based Gait Recognition Methods

Though we have not found any articles on gait recognition based on fusion with camera and radar sensors, there have been some pioneering works on multisensor fusion in the gait recognition community. Zou et al. [6] creatively proposed a robust gait recognition method for cross-walking-condition by fusing the intertial and RGBD sensors. They achieved promising performance with the help of a support vector machine (SVM) classifier. Kumar et al. [47] proposed a multimodal gait recognition method and extract the gait information of camera data and intertial data by using the 3-D CNNs and LSTMs, respectively. They achieved better performance using deep learning methods. Zhao et al. [7] proposed a new neural network architecture by extracting temporal features and spatial features and fusing the results. They focus on neurodegenerative diseases and person identity classification.

However, these methods use wearable sensors as one of the sources, which requires personnel cooperation and carry-on. Meanwhile, these methods also did not explore gait recognition under cross-view conditions. In this article, we utilize the camera and radar sensors to capture the gait information since camera and radar are two popular nonwearable sensors that can capture gait information from long distances without human cooperation. Besides that, radar can further capture micro-motion pattern information, such as swings of human arms, legs, feet, and other parts, which is very conductive to gait recognition. We also explore the gait recognition under both cross-view and cross-walking-condition.

III. Proposed Method

A. System Overview

Fig. 2 gives the overview of our gait-based human identification system with multisensor fusion. When a person enters the observation area with different views and walking conditions, we use an RGB camera (BASLER acA1600-60gc) to collect optical data and use a millimeter-wave radar (Texas Instruments 77-GHZ AWR1843) to collect radar data. The raw optical data is presented in the form of video, which can portray the information of human appearance very well. We use the deeplabV3+ [48] which is pretrained with a large-scale human body segmentation data set [49], [50] to detect

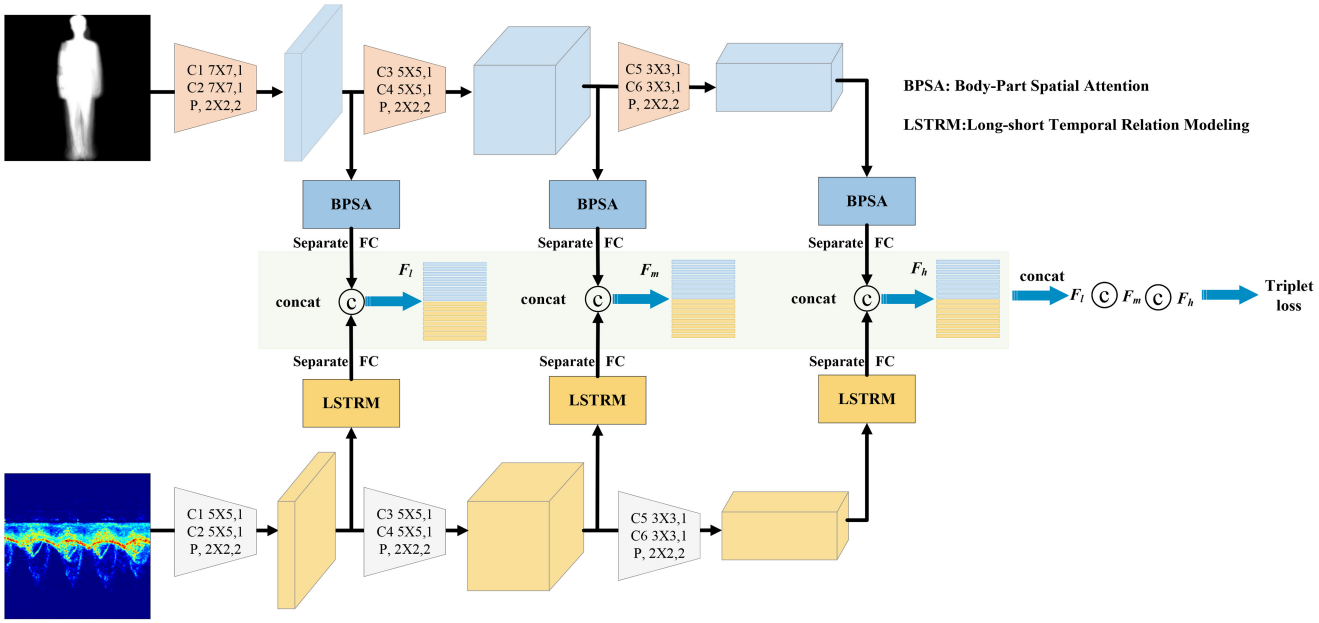


Fig. 3. Overall framework, which consists of two-stream CNNs, BPSA module, LSTRM module, MSF, and metric learning. Triplet loss is applied to operate the metric learning.

and segment the person in raw video. After person detection and segmentation, we can obtain the silhouette sequence, which is further compressed to obtain GEIs [18]. Millimeter-wave radar emits linear frequency modulated continuous wave (LFMCW) and receives the echo signal reflected from the human body. The time-Doppler spectrograms of the echo signal can be obtained by time-frequency analysis, which reflect the transformed relationship between micro-Doppler frequency and amplitude of human parts with time. The specific process includes: 1) mixing the echo signal with the transmit signal to obtain the intermediate frequency (IF) signal; 2) doing the fast Fourier transform along the range dimension (Range FFT) on the IF signal to obtain the time-range map; 3) performing static clutter removal (SCR) in the time-range map; and 4) selecting the range unit where the person is located in the time-range map to do the short-time Fourier transform (STFT).

After data processing, proposed method then extracts the gait features from GEIs and time-Doppler spectrograms and fuses them to obtain the richer and more robust features for human identification. In our article, gait recognition is also regarded as a retrieval task. In the training process, we end-to-end optimize feature discrepancy of intraidentity and interidentity with deep CNNs and metric learning. In the test process, given the samples to be tested, i.e., probe samples, gait recognition system needs to find the best match, with the nearest classifier, based on computed similarities (distances) from the gallery.

The proposed method can capture and recognize the gait from long distances without human cooperation. In the following, we will elaborate on the proposed robust gait recognition with multisensor fusion.

B. Robust Gait Recognition With Multisensor Fusion

The overall architecture of the proposed robust gait recognition with multisensor fusion is shown in Fig. 3, which consists of two-stream CNNs, BPSA module, LSTRM module, multiscale fusion (MSF), and metric learning. A batch of B GEIs samples and a batch of B time-Doppler spectrogram samples are fed into the two-stream CNNs as input, which is represented as $G \in \mathbb{R}^{B\times 1\times H\times W}$ and $Spe \in \mathbb{R}^{B\times 3\times D\times T}$, respectively. H and W denote the height and width of GEIs while D and T denote the Doppler and time dimension of time-Doppler spectrograms. First, G and Spe are passed through a two-stream CNNs with 3 layers to extract multiscale features, which are denoted as low-level gait features: $F_{GL} \in \mathbb{R}^{B\times C_L\times H/2\times W/2}$ and $F_{TL} \in \mathbb{R}^{B\times C_L\times D/2\times T/2}$, mid-level gait features: $F_{GM} \in \mathbb{R}^{B\times C_M\times H/4\times W/4}$ and $F_{TM} \in \mathbb{R}^{B\times C_M\times D/4\times T/4}$, and high-level gait features: $F_{GH} \in \mathbb{R}^{B\times C_H\times H/8\times W/8}$ and $F_{TH} \in \mathbb{R}^{B\times C_H\times D/8\times T/8}$, where C_L, C_M, and C_H denote the number of feature channel at different levels. For an input x, each layer can be represented as

$$\text{out} = \text{pool}(f(W_2(f(W_1(x))))) \tag{1}$$

where W_i denotes the ith convolution operations, f denotes the ReLU activation function, and pool denotes the pooling operation. The pooling operation can somewhat alleviate the sensitivity to spatial displacement.

Afterward, we implement a BPSA module to obtain more discriminative body part representations of GEIs and an LSTRM module to obtain the local and global micro-motion representation of time-Doppler spectrograms. Then the gait features of GEIs and time-Doppler spectrograms are fused at the multiscale feature space. Finally, the multiscale fused features are concatenated as outputs for triplet loss optimization.

1) Body-Part Spatial Attention: For camera GEIs data, variations in viewpoint or walking conditions generally result in changes in body appearance. Gait recognition methods that directly learn the overall representation of the human body tend to be sensitive to changes in appearance, which is fragile for human identification under complex covariations.

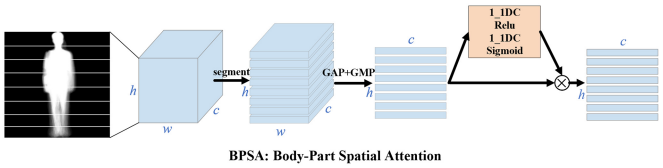


Fig. 4. Structure of BPSA module, which consists of body-part segmentation and attention mechanism.

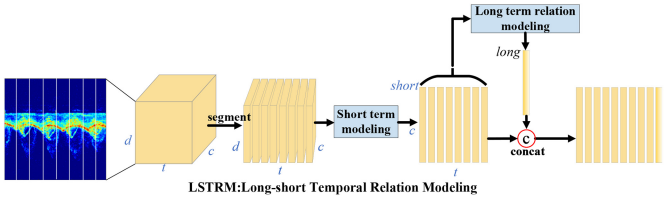


Fig. 5. Structure of LSTRM module, which consists of short-term modeling and long-term relation modeling.

Unlike learning the global representation of the human body, body regions generally maintain different contributions toward person reidentification/gait recognition under cross-view and cross-walking-condition [14], [51].

We aim to learn the importance of different body parts (regions) and improve the robustness of gait-based human identification against variations in viewpoint and walking conditions. The proposed BPSA is shown in Fig. 4. First, we segment the feature map $F \in \mathbb{R}^{B\times C\times H_1\times W_1}$ into N parts, which is denoted as $P \in \mathbb{R}^{B\times C\times (H_1/N)\times W_1\times N}$, and each part represents a region of the human body. Then global pooling operation is applied to aggregate spatial information of body regions to get a compact part descriptor $P_{gp} \in \mathbb{R}^{B\times C\times N}$. Since global max pooling can perceive the discriminative information in the region and global average pooling (GAP) can perceive the overall information of the region, we applied both GMP and GAP. The operation can be denoted as

$$P_{gp} = \mathrm{GMP}(P) + \mathrm{GAP}(P). \tag{2}$$

For persons of cross-views and cross-walking-condition, different parts of the body are divided according to regions, and different parts are of varying importance for gait recognition performance under complex covariates. To learn the importance of different body parts, we proposed an attention mechanism to obtain a more discriminative body representation. The part descriptor is forwarded to the attention mechanism to produce the attention map. The attention mechanism can be formulated as

$$I = \sigma\left(W_2\delta\left(W_1 P_{gp}\right)\right) \tag{3}$$

where σ refers to the sigmoid function, δ refers to the LeakyReLU function, $W_1 \in \mathbb{R}^{(C/r)\times C}$, $W_2 \in \mathbb{R}^{1\times (C/r)}$, and r is the reduction ratio. To reduce the complexity, we use two sets of 1-D convolutions to implement.

$I \in \mathbb{R}^{1\times N}$ reflects the importance of each body part, and then P_{gp} is multiplied by I to get the enhanced feature map

$$F_a = I \otimes p_{gp} \tag{4}$$

where $\otimes$ represents elementwise multiplication. During the multiplication, the importance values are broadcasted accordingly along with the body parts.

2) Long-Short Temporal Relation Modeling: Time-Doppler spectrograms reflect the transformation of micro-Doppler frequencies and amplitudes in various parts of the body over time. Although CNNs have achieved guaranteed performance with their powerful ability to automatically learn discriminative feature representations, convolutional operations process a local neighborhood and cannot effectively capture long-term dependencies, which is not conducive to the learning of human micro-motion patterns.

To alleviate the above issue, we propose an LSTRM module to obtain the local and global micro-motion representation of time-Doppler spectrograms. The proposed LSTRM is shown in Fig. 5. We first split the feature map $F \in \mathbb{R}^{B\times C\times D_1\times T_1}$ along with the time dimension to get the M short-term strips, which can be denoted as $S \in \mathbb{R}^{B\times C\times D\times (T/M)\times M}$. The short-term strips reflect the local micro-motion. To model the short-term temporal information, the global pooling operation, which includes global max pooling and GAP, is applied to aggregate information of short-term strips to get a compact short-term descriptor, which is denoted as $S_{gp} \in \mathbb{R}^{B\times C\times M}$. Then a long-term relation modeling mechanism is proposed to obtain the global micro-pattern: the multilayer perception machine (MLP) and Sigmoid functions are used to evaluate each short-time feature, and then all the short-term features are weighted sum to obtain the long-term features reflecting the relationship of the short-time features. In short, the computation process can be formulated as

$$\begin{aligned} L &= \frac{\sum_{m=1}^{M}\sigma\left(MLP\left(S_{gp}\right)\right)\cdot S_{gp}}{\sum_{m=1}^{M}\sigma\left(MLP\left(S_{gp}\right)\right)} \\ &= \frac{\sum_{m=1}^{M}\sigma\left(W_2\left(\delta\left(W_1\left(S_{gp}\right)\right)\right)\right)\cdot S_{gp}}{\sum_{m=1}^{M}\sigma\left(W_2\left(\delta\left(W_1\left(S_{gp}\right)\right)\right)\right)}. \end{aligned} \tag{5}$$

where σ refers to the sigmoid function, δ refers to the LeakyReLU function, $W_1 \in \mathbb{R}^{(c/r)\times c}$ and $W_2 \in \mathbb{R}^{c\times (c/r)}$.

The short-term features can reflect the local micro-pattern of the human body while the long-term features can reflect the global micro-pattern of the human body. We finally integrate the local and global micro-pattern by using the concatenation operation.

3) Multiscale Fusion: After the BPSA and LSTRM with two-stream CNNs, we can obtain the low-, mid-, and high-level features of GEIs and time-Doppler spectrograms. Different layers of CNNs have different receptive fields and features of different levels focus on different information for gait recognition. Lower-level features focus more on fine-grained gait information while higher-level features focus more on coarse-fined and semantic gait information. Therefore, we fuse the gait representations of GEIs and time-Doppler spectrograms at the multiscale feature space. Specifically, the features after BPSA and LSTRM first are inputted into the separate fully connected layer (FC) to obtain the final features for fusion. The separate FC uses separate FCs to process the gait features of the two modalities, respectively, which can better preserve the complementary information.

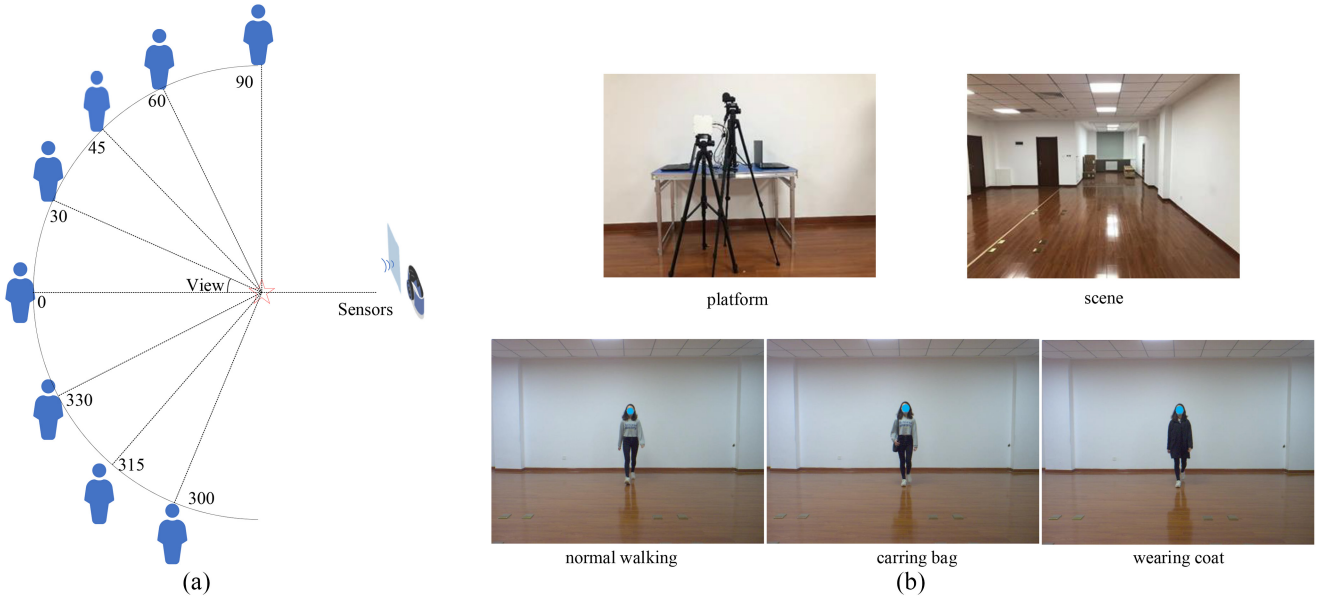


Fig. 6. Walking views, platform, real scene of data collection, and examples of three walking conditions. (a) Eight walking views. (b) Platform, scene, and examples of three walking conditions.

Denote the low-, mid-, and high-level features of GEIs and time-Doppler spectrograms after FC as $F_{GL}, F_{GM}, F_{GH} \in \mathbb{R}^{B\times N\times \text{Hidd}}$ and $F_{TL}, F_{TM}, F_{TH} \in \mathbb{R}^{B\times \text{Hidd}\times(M+1)}$, where Hidd denotes the number of hidden layer neurons of FC. Then we use the concatenation operation to fuse the multiscale features of GEIs and spectrograms. The fusion process is shown in Fig. 3, which can be described as follows:

$$F_{\text{fusion}} = \left(F_{GL} \copyright F_{TL}^{T}\right) \copyright \left(F_{GM} \copyright F_{TM}^{T}\right) \copyright \left(F_{GH} \copyright F_{TH}^{T}\right) \quad (6)$$

where © denotes the concatenation operation and $F_{\text{fusion}} \in \mathbb{R}^{B\times 3(N+M+1)\times \text{Hidd}}$.

Note that we use the simple concatenation to fuse the features of GEIs and time-Doppler spectrograms. The reason is that the appearance and micro-motion are completely two different but highly complementary information for gait recognition. Our fusion way aims to reserve the complementary information of the two modalities to the maximum extent. We designed the different fusion technologies in the experiment to explore this issue, including data fusion, decision fusion, attention-based feature fusion, first-order feature fusion, for instance, elementwise sum, and second-order fusion, for instance, transformer. The result and analysis will be elaborated on in the next section.

Training Loss: Batch all (BA) triplet loss [52] is applied as the metric learning loss function. The corresponding output fusion features of different samples are used to compute the triplet loss. BA triplet loss is described as follows:

$$\mathcal{L}_{\text{BA}}(\theta; X) = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{K}}^{\text{all anchors}} \overbrace{\sum_{\substack{p=1\\ p\neq a}}^{K}}^{\text{all pos.}} \overbrace{\sum_{\substack{j=1\\ j\neq i}}^{P}\sum_{n=1}^{K}}^{\text{all neg.}} \left(\max\left(\text{margin} - d_{\text{neg}}^{2} + d_{\text{pos}}^{2}, 0\right)\right) \quad (7)$$

where P denotes the number of sampling person identities, K denotes the number of sampling samples of each person, d_{neg}^{2} denotes the distance between features of negative pairs, d_{pos}^{2} denotes the distance between features of positive pairs, and margin is a threshold set by experience. BA triplet loss aims to maximize the interidentity distance and minimize the intraidentity distance.

Testing: Given the samples to be tested, i.e., probe samples, the gait recognition system needs to find the best match, with the nearest classifier, based on computed similarities (distances) from the gallery. Specifically, the probe set and gallery set are first inputted into the network to obtain the multiscale fused feature descriptors. Then we calculate the Euclidean distance of the probe feature descriptor with every gallery feature descriptor. Finally, each probe sample is assigned the identity using the nearest distance (Rank1 score).

IV. Experiments and Analysis

A. Data Set Collection and Description

Following the data collection of the human identity system shown in Fig. 2, we use the RGB camera and millimeter-wave radar to collect gait data from 121 subjects with eight views and three walking conditions, which takes more than one month for overall data collection. Fig. 6(a) shows the schematic of the eight walking views, and Fig. 6(b) shows examples of three walking conditions, platform, and scene of data collection. The data set consisted of 72 men and 49 women, aged 21 to 25 years, weighing between 42 and 100 kg, and measuring 155–187 cm in height. The camera and radar sensors simultaneously record 2.4 s of gait information for each walk. The frame rate of the camera and radar sensors is both set as 50 frame/s.

There are 80 pairs (GEIs and time-Doppler spectrograms) per subject, i.e., 8 views (0, 30, 45, 60, 90, 300, 315, 330 degree) and 10 groups for each view. Among the ten groups, there are six groups collected under normal walking (NM)

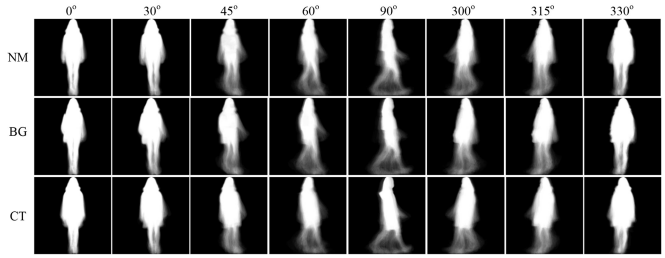


Fig. 7. Some examples of GEIs in our data set. Left to right: different views. Top to bottom: different walking conditions.

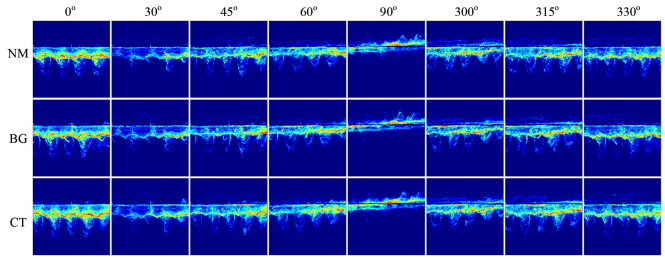


Fig. 8. Some examples of time-Doppler spectrograms in our data set. Left to right: different views. Top to bottom: different walking conditions.

conditions, four of them are set in the gallery, which is denoted as NM#1-4, and the remaining two groups are set in the probe, which is denoted as NM#5-6. There are two groups collected under carrying bags (BG) and they are set to probe, which is denoted as BG#1-2. The last two groups are collected under wearing coats (CT) and they also are set in the probe, which is denoted as CT#1-2. Some examples of GEIs and time-Doppler spectrograms are shown in Figs. 7 and 8. In our experiments, the first 74 persons are used to train the network and the rest 47 persons are used to test.

B. Implementation Details

In our experiments, all GEIs are resized to $H \times W = 128 \times 88$ and time-Doppler spectrograms are resized to $D \times T = 88 \times 128$. We choose the Adam as optimizer. C_L, C_M, and C_H in two-stream CNNs are set as 32, 64, 128, respectively. The segmentation parts of BPSA and LSTRM are set as 16 and 15, respectively. The number of hidden layer neurons of FC is set as 256. The margin, P and K of BA triplet loss are set as 0.2, 4, and 2, respectively. The learning rate is 0.0001 and we train the network for 300k iterations. Our method is implemented by Pytorch with NVIDIA GeForce GTX 2080Ti GPU.

C. Evaluation

We evaluate gait recognition performance under complex covariate conditions, namely, cross-view and cross-walking-conditions. During the training process, we do have access to all walking conditions and views. However, during the test process, we report the performance of cross-view and cross-walking-conditions, i.e., given the specific probe, the identical view or walking conditions in the gallery are excluded. We report the rank-1 accuracy.

Cross-View: All the results are averaged on the seven gallery views and the identical views are excluded. For example, the

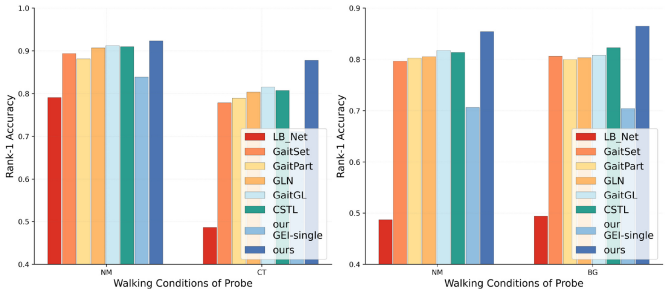


Fig. 9. Cross-view recognition accuracies of radar-based methods under cross-walking-conditions, where "NM," "BG," and "CT" represent the walking condition of probe samples. (a) BG#1-2 are set as gallery samples. (b) CT#1-2 are set as gallery samples.

performance of probe view 0° is averaged on other seven gallery views and gallery view 0° is excluded. For the accuracy of each probe view, the above computing process can be formulated as

$$\mathrm{Acc}_i = \frac{1}{7} \sum_{\substack{j=1 \\ j \neq i}}^{7} \mathrm{cross}(i, j) \tag{8}$$

where j denotes the index of gallery view, $\mathrm{cross}(i, j)$ represents the rank1 accuracy when the probe view index is i and the gallery view index is j.

Cross-Walking-Condition: The accuracy of the cross-walking-condition is reported when the walking conditions between the probe and gallery are different. For example, when the walking condition of the gallery is NM, we report the performance when the condition of the probe is carrying the bag.

In our experiments, NM#1-4 are set as the gallery samples. When the probe is NM#5-6, we report the performance of cross-view. When the probe is BG#1-2 or CT#1-2, we report the performance both of cross-view and cross-walking-condition.

D. Performance

1) Compared With Camera-Based Methods: In the gait recognition literature, camera-based methods dominate in terms of performance. To verify the effectiveness of our method, in this section, we compare our proposed method with several state-of-the-art camera-based methods under cross-view and cross-walking-conditions in Table I and Fig. 9, where "LB_Net" denotes the GEIs-based method proposed in [19], "GaitSet," "GaitPart," "GLN," "GaitGL," and "CSTL" are the sequence-based methods proposed in [14], [15], [16], [28], and [29], respectively. "our GEI-single" denotes our GEIs-branch, and "ours" denotes our robust gait recognition method with camera and radar fusion. Table I reports the cross-view accuracy of probe samples under different walking conditions when the gallery samples are under NM conditions. Fig. 9 reports the cross-view accuracy of probe samples under different walking conditions when the gallery samples are under carrying bag and wearing coat conditions.

TABLE I
COMPARISON OF THE OVERALL PERFORMANCE, EXCLUDING IDENTICAL-VIEW CASES

Gallery NM#1-4		Gallery (8 views)								Mean
Probe		0	30	45	60	90	300	315	330	
NM#5-6	LB_Net	73.17	81.65	85.62	80.36	65.32	81.23	83.24	78.36	78.619
	GaitSet	88.09	91.07	97.47	96.87	80.95	94.05	96.58	88.39	91.685
	GaitPart	89.21	93.31	98.18	95.74	78.12	94.07	96.05	90.12	91.851
	GLN	87.23	91.03	98.18	95.14	82.37	94.53	96.81	91.95	92.155
	GaitGL	88.75	91.79	98.63	**97.11**	84.80	95.29	97.57	93.92	93.484
	CSTL	88.36	91.55	97.63	96.57	84.41	95.20	97.02	93.68	93.054
	our GEI_single	85.32	88.73	95.1	87.95	72.27	88.73	91.99	87.49	87.198
	ours	**93.95**	**94.1**	**98.76**	96.90	**90.22**	**97.36**	**97.83**	**94.41**	**95.439**
BG#1-2	LB_Net	68.32	72.13	75.64	71.11	56.24	71.89	74.28	70.01	66.202
	GaitSet	84.82	86.76	94.64	93.9	79.91	93.6	94.49	85.42	89.193
	GaitPart	86.78	89.82	96.05	**95.29**	76.44	93.01	93.62	86.63	89.704
	GLN	84.35	88.14	96.35	92.70	79.18	91.79	94.68	88.45	89.456
	GaitGL	86.99	89.27	96.26	94.14	79.85	91.10	95.66	88.97	90.281
	CSTL	88.67	91.86	96.11	92.53	80.00	91.71	95.35	**89.85**	90.851
	our GEI_single	82.92	84.94	88.51	83.23	68.79	83.85	87.42	76.86	82.066
	ours	**93.48**	**92.08**	95.96	92.55	**83.23**	**95.65**	**96.27**	88.20	**92.178**
CT#1-2	LB_Net	50.16	53.34	64.63	50.23	32.61	44.22	47.16	42.13	48.06
	GaitSet	70.24	73.96	86.46	84.97	70.54	77.68	78.87	72.92	76.954
	GaitPart	73.86	83.28	87.54	83.98	69	82.83	83.13	75.68	79.902
	GLN	69.60	74.47	84.95	81.15	70.82	82.22	86.17	76.60	78.248
	GaitGL	72.34	79.79	86.63	82.98	74.47	81.31	85.87	78.27	80.205
	CSTL	74.68	81.37	86.84	82.13	73.32	81.07	86.99	81.67	81.009
	our GEI_single	68.15	73.12	79.64	73.43	60.85	74.36	74.69	67.98	71.527
	ours	**85.56**	**88.04**	**90.84**	**89.13**	**78.73**	**87.89**	**90.53**	**86.49**	**87.151**

As shown in Table I and Fig. 9, our method significantly outperforms other methods in various cross-view and cross-walking-condition, especially, under CT condition, demonstrating the robustness and effectiveness of our method. As shown in Table I, in terms of the relationship between views and accuracy, we observed that the performance of 90° is almost the worst in all cases and methods. The possible reason is that the 90° view of the pedestrian has changed considerably in both appearance and motion pattern. The performance of 45° is almost the best in all cases and methods. The possible reason is that the 45° has a minimal difference from the other views in our data set. Compared to the GEI-single, our method improves the mean accuracy of various cases by fusing the gait features of time-Doppler spectrograms, which improves the cross-view accuracy of 8.241%, 10.112%, and 15.624% under NM, BG, and CT, respectively. This observation also verifies the huge potential of camera and radar sensor fusion under complex covariations. It is also noted that the performance of our method drops less when the walking condition and views are both changed. Taking the GaitGL as an example, the mean cross-view accuracy drops about 13% (from 93.484% to 80.205%) when the walking condition changes from the NM to the CT while the mean accuracy of our method drops about 8% (from 95.439% to 87.151%). The reason is that our method extracts more robust gait features by fusing the discriminant appearance representation and micro-motion. In summary, comparison results with other approaches in Table I and Fig. 9 demonstrate that the gait recognition accuracy and robustness can be significantly improved by fusing camera and radar sensors.

2) Compared With Radar-Based Methods: Due to the lack of the gait recognition method based on the retrieval task for time-Doppler spectrograms, we reproduce several classical gait recognition methods for classification-based tasks and replace the classification loss with a BA triplet loss to perform the retrieval task. Fig. 10 gives the results of different methods, where "model1," "model2," "model3," MBCNN, and AttentionGait represent the CNN-based method proposed in [38], [39], [40], [53], and [54], respectively, "TCN" represents the method in [41] based on temporal CNNs, our radar-single denotes our radar-branch, and ours denotes our robust gait recognition method with camera and radar fusion. As shown in Fig. 10, most CNN-based methods alone do not yield satisfactory results since these methods do not fully utilize temporal information. TCN achieved performance improvements by using 1-D convolutions to model the short temporal information. Due to the ability of multiscale temporal feature extraction, MBCNN achieves better performance than other comparison methods. The proposed method with LSTRM achieves the best performance since finer long and short temporal relationships are modeled. Compared to the radar-single, our method significantly improves the mean accuracy of various cases by fusing the gait features of GEIs.

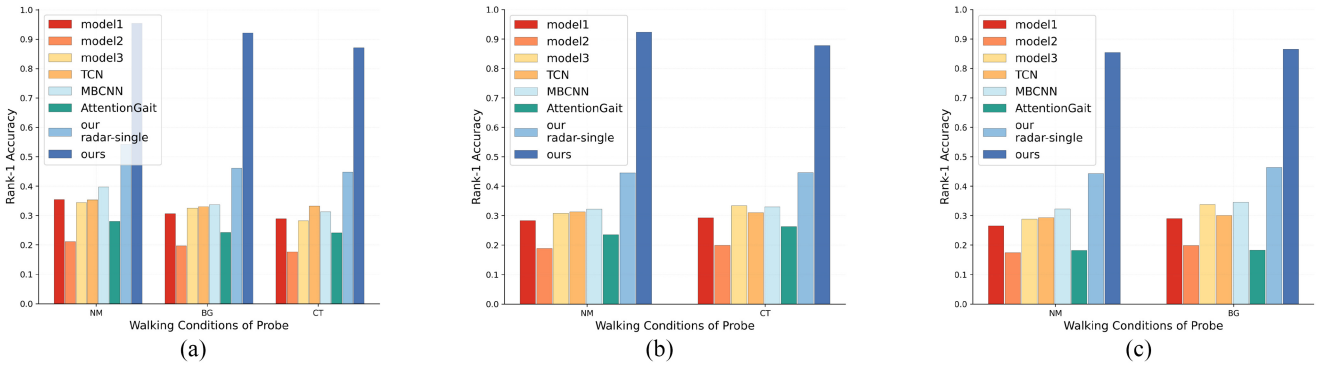


Fig. 10. Cross-view recognition accuracies of radar-based methods under cross-walking-conditions, where "NM," "BG," and "CT" represent the walking-condition of probe samples. (a) NM#5-6 are set as the gallery samples. (b) BG#1-2 are set as the gallery samples. (c) CT#1-2 are set as the gallery samples.

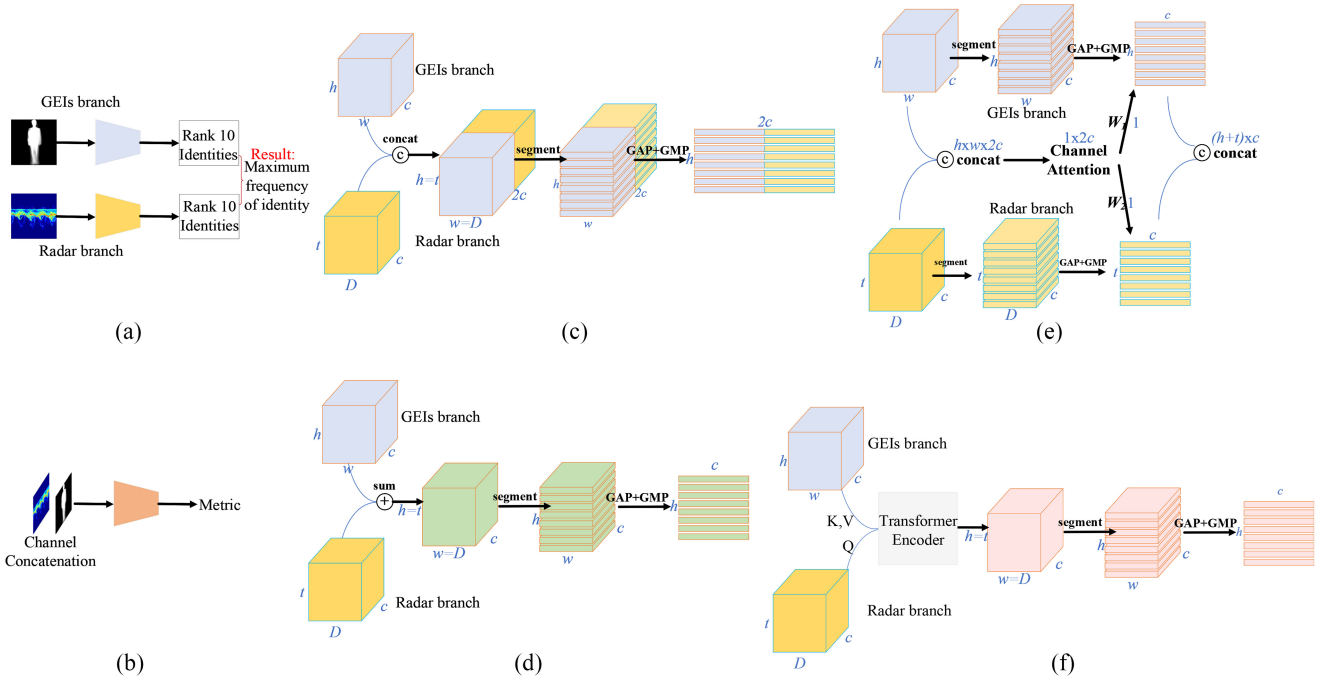


Fig. 11. Structures of different designed fusion technologies, where h and w denote the height and width of GEIs while D and T denote the Doppler and time dimension of time-Doppler spectrograms, c denotes the number of channel numbers. (c) and (d) give the single-scale illustrations for simplicity. (a) Decision fusion. (b) Data fusion. (c) Channel concatenation. (d) Elementwise sum. (e) Attention-based reweight. (f) Transformer-based fusion.

E. Impact of Fusion Ways

In this section, we aim to analyze the impact of fusion ways. We compare the other nine fusion technologies in the experiment, including data fusion, decision fusion, attention-based feature fusion, first-order feature fusion, and second-order fusion, for instance, transformer. The details and structure of the first six comparison fusion technologies are described in Fig. 11 and nine fusion technologies are described as follows.

1) *Decision Fusion:* We use a voting mechanism to fuse the result of GEIs-branch and radar-branch. Specifically, for each branch, we extract the rank10 IDs and combined the IDs of both branches. We select the identity with the highest frequency of occurrence as the final prediction. In short, we refer to the method as "Decision." Note that for decision fusion, we do not evaluate with rank-1 accuracy.
2) *Data Fusion:* We fuse the gait information at the data level by using the concatenation function to connect the two types of data along the channel dimension. Then the fused data flows to the single-stream 3-layers CNNs with feature segmentation and triplet loss. In short, we refer the method as "Data."
3) *Multiscale Channel Concatenation:* We fuse the gait information at the multiscale feature level by using the concatenation function to connect the two types of features along the channel dimension. The multiscale fused features are then further connected after feature segmentation to obtain the final fused features for triplet loss optimization. In short, we refer to the method as "Channel."
4) *Multiscale Elementwise Sum:* We fuse the gait information at the multiscale feature level by using the

elementwise sum function to connect the two types of features. The multiscale fused features are then further connected after feature segmentation to obtain the final fused features for triplet loss optimization. In short, we refer to the method as "Sum."

5) *Attention-Based Reweight:* We first use the concatenation function to connect the two types of features along the channel dimension. Then SE attention mechanism is applied to learn the interactive relationships between the two features. Denote the channels of each feature map is c dimension, after the attention mechanism, we can get the weights vector with $2c$ dimension. For the corresponding c dimension of each feature map, we aggregate the c dimension weights into one weight. The weights are multiplied by the corresponding branch with feature segmentation. Finally, the reweighted branches are fused with concatenation to obtain the final features for triplet loss optimization. In short, we refer to the method as "Attention."
6) *Transformer-Based Fusion:* We fuse the gait information at the last convolutional layer by using a transformer [55] to connect the two types of features. The final fused features are then obtained after feature segmentation for triplet loss optimization. In short, we refer to the method as "Transformer."
7) *GRIFNet [56]:* GRIFNet combines the features of radar point cloud data and camera data by the convolutional Mixture of Experts (MoE) with convolutional layers and sigmoid layers so that MoE explicitly assigns weights to the features of two modalities.
8) *Transfuer [57]:* TransFuser proposes a mechanism to integrate image and LiDAR representations using self-attention and uses two-branches transformer modules at multiple resolutions to fuse feature maps of LiDAR point cloud data and camera data.
9) *MCLNet [58]:* MCLNet is designed to learn modality-invariant features between camera and infrared data by simultaneously minimizing intermodality discrepancy while maximizing cross-modality similarity among instances.

Table II shows the results of different fusion technologies. Several conclusions are summarized as follows.

1) Simple fusion ways, namely, decision fusion and data fusion, can also achieve good performance, which demonstrates highly complementary gait information can be captured by the camera and radar sensors. This result also verifies the motivation of our method from the side.
2) One-order fusion ways, i.e., Channel, Sum, and our method achieve better performance than second-order fusion way, i.e., Transformer, "Transfuser," and "MCLNet." Transformer-based methods shine in many fields of multimodality, as the transformer can capture higher-order similarities between modalities. Transfuser loses less modal information than Transformer and achieves better recognition performance. However, in our scenarios, the GEIs modality mainly reflects the appearance information of gait while the modality of time-Doppler spectrograms mainly reflects the micro-motion information of gait. The appearance and micro-motion are completely two different but highly complementary information. Therefore, learning the similarities between these two types of information may be counterproductive. MCLNet is designed to learn modality-invariant features between camera and radar data. This fusion strategy losses the complementarity of two types of gait information to a certain extent, making the fusion way of MCLNet poor performance.
3) Our method, which uses the concatenation operation to fuse two types of gait information along the space and temporal dimensions, achieves the best performance. The fusion way of "multiscale channel concatenation" uses the concatenation operation to fuse two types of gait information along the channel dimension. Since the fusion features after channel concatenation will be integrated by FC, the complementarity of two types of gait information may be lost to a certain extent, making the fusion way of multiscale channel concatenation slightly poor performance. Compared to the multiscale channel concatenation, two types of gait features with our fusion way can be processed separately by the separate FC, which preserves the maximum complementarity of the two modalities. The fusion way of "attention-based reweight" and "GRIFNet" attempt to learn the relationship between the two modalities with the channel attention mechanism and assign weights to the two modalities. Since learning the relations between these two types of information may be counterproductive, the performance has dropped compared with our method, despite it also uses the concatenation operation to fuse two weighted modalities along the space and temporal dimensions.

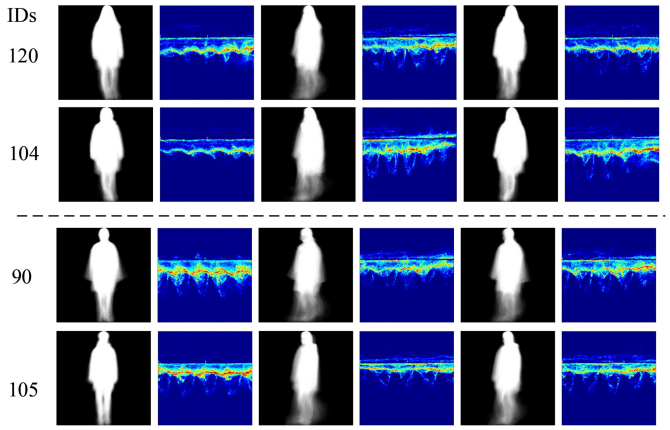


Fig. 12. Examples where the GEIs are difficult to distinguish but the time-Doppler spectrograms can be distinguished. The numbers on the left column represent the index of identities in our test data. We can see that the GEIs of ID_120 and ID_104 are similar, but the time-Doppler spectrograms of ID_120 and ID_104 can be distinguished; the GEIs of ID_90 and ID_105 are similar, but the time-Doppler spectrograms of ID_90 and ID_105 can be distinguished.

We further explore the complementarity of the two modalities. Fig. 12 shows some examples where the GEIs are difficult to distinguish but the time-Doppler spectrograms can

TABLE II
COMPARISON OF THE OVERALL PERFORMANCE OF DIFFERENT FUSION TECHNOLOGIES, EXCLUDING IDENTICAL-VIEW CASES

Gallery NM#1-4		Gallery (8 views)								Mean
Probe		0	30	45	60	90	300	315	330	
NM#5-6	GEI_single	85.32	88.73	95.1	87.95	72.27	88.73	91.99	87.49	87.198
	Radar_single	52.33	56.99	57.45	58.07	39.75	52.95	59.63	56.68	54.232
	Decision	86.18	92.86	95.96	94.1	81.52	93.63	92.86	91.15	91.032
	Data	90.84	91.93	96.12	93.32	80.43	93.01	95.5	90.68	91.479
	Channel	91.31	95.03	97.83	96.12	87.11	95.34	96.12	93.79	94.08
	Sum	94.1	93.17	98.14	95.65	85.87	94.57	96.9	92.24	93.828
	Attention	91.08	92.48	95.74	93.57	83.32	93.26	94.96	93.25	92.207
	Transformer	79.81	83.23	89.29	86.34	66.46	82.45	87.58	83.23	82.298
	GRIFNet	86.78	86.42	92.84	91.91	79.33	92.06	92.84	91.44	89.577
	Transfuer	87.11	91.00	94.26	**90.99**	80.28	91.46	94.41	90.37	89.985
	MCLNet	91.62	93.32	96.12	93.94	84.78	94.56	97.20	91.92	92.934
	ours	**93.95**	**94.1**	**98.76**	**96.9**	**90.22**	**97.36**	**97.83**	**94.41**	**95.439**
BG#1-2	GEI_single	82.92	84.94	88.51	83.23	68.79	83.85	87.42	76.86	82.066
	Radar_single	43.63	41.161	43.17	46.43	37.11	52.79	56.37	47.98	46.137
	Decision	81.99	86.34	91.93	89.13	75	90.53	93.48	85.25	86.704
	Data	86.42	88.44	93.88	85.95	76.64	88.91	93.25	85.49	87.373
	Channel	87.89	91.46	94.72	90.68	81.68	91.31	94.56	87.42	89.966
	Sum	89.75	88.51	94.57	88.66	79.97	91.3	93.94	88.36	89.383
	Attention	88.2	90.99	93.32	88.2	81.37	92.24	92.24	86.03	89.073
	Transformer	74.38	74.22	83.54	81.05	63.66	77.17	83.7	70.96	76.087
	GRIFNet	86.01	83.68	90.04	86.47	75.29	84.61	89.42	84.14	84.958
	Transfuer	79.66	83.07	90.53	86.49	73.14	87.27	92.55	83.85	84.568
	MCLNet	86.65	88.51	92.24	88.35	78.57	91.46	93.79	88.66	88.529
	ours	**93.48**	**92.08**	**95.96**	**92.55**	**83.23**	**95.65**	**96.27**	**88.2**	**92.178**
CT#1-2	GEI_single	68.15	73.12	79.64	73.43	60.85	74.36	74.69	67.98	71.527
	Radar_single	46.89	42.24	49.38	45.65	34.32	44.87	50.31	44.56	44.778
	Decision	75.78	77.64	83.54	83.7	71.89	80.55	82.14	79.5	79.329
	Data	79.81	79.81	86.49	82.61	69.25	78.88	83.07	80.75	80.085
	Channel	82.52	83.92	89.97	88.27	76.93	84.39	87.8	79.42	84.152
	Sum	84.01	84.94	88.82	86.49	75.31	82.14	84.01	76.55	82.784
	Attention	80.28	84.94	87.27	84.94	72.67	80.59	85.09	78.73	81.813
	Transformer	60.87	66.77	73.45	65.37	57.14	65.22	65.68	60.4	64.363
	GRIFNet	75.14	79.33	81.82	78.55	70.01	73.27	80.10	78.09	77.039
	Transfuer	68.79	74.07	80.74	80.43	69.41	77.33	78.26	71.58	75.078
	MCLNet	80.59	82.76	86.18	81.83	71.47	83.70	84.78	81.68	81.657
	ours	**85.56**	**88.04**	**90.84**	**89.13**	**78.73**	**87.89**	**90.53**	**86.49**	**87.151**

be distinguished. In contrast, Fig. 13 shows some examples where the time-Doppler spectrograms are difficult to be distinguished but the GEIs can be distinguished. The GEIs can well reflect the appearance information of gait but lack the fine-grained information of gait micro-motion patterns while the time-Doppler spectrograms can well portray the micro-motion patterns of gait but lack the appearance information of gait. Thus, for some individuals, as shown in Fig. 12, they look similar but the fine-grained micro-motion patterns can be distinguished. Similarly, as shown in Fig. 13, for some individuals, their micro-motor patterns may be similar, but their appearance (e.g., the obesity levels of the two groups of individuals in Fig. 13) is well differentiated. These showcases can further verify why even the simplest of decision fusions can work well. The gait information captured by camera and radar can alleviate the different kinds of errors and the sensitivity to complex covariate conditions introduced by a single type of sensor.

F. Model Analysis

1) Ablation Study: To comprehensively investigate the effectiveness of different modules on gait recognition performance, we conduct the ablation study by evaluating variants of the proposed method. The ablation study results are shown in Table III, where BPSA, LSTRM, SSF, and MSF

TABLE III
ABLATION STUDY

	BPSA	LSTRM	SSF	MSF	NM	BG	CT	Mean
GEI-single		-	-	-	84.957	80.59	68.906	78.151
	✓	-	-	-	87.198	82.066	71.527	80.264
Radar-single	-		-	-	50.001	43.103	41.75	44.951
	-	✓	-	-	54.232	46.137	44.778	48.382
GEIs+Radar			✓		91.994	88.005	80.512	86.837
	✓		✓		92.449	89.033	81.172	87.551
		✓	✓		93.42	89.644	83.851	88.971
	✓			✓	95.031	91.538	86.278	90.949
		✓		✓	95.303	91.77	86.51	91.194
	✓	✓	✓		93.692	91.257	84.093	89.681
	✓	✓		✓	**95.439**	**92.178**	**87.151**	**91.589**

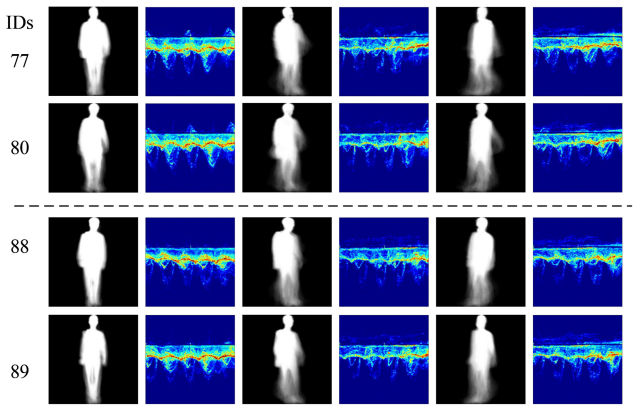


Fig. 13. Examples where the time-Doppler spectrograms are difficult to distinguish but the GEIs can be distinguished. The numbers on the left column represent the index of identities in our test data. We can see that the time-Doppler spectrograms of ID_77 and ID_80 are similar, but the GEIs of ID_77 and ID_80 can be distinguished; the time-Doppler spectrograms of ID_88 and ID_89 are similar, but the GEIs of ID_88 and ID_89 can be distinguished.

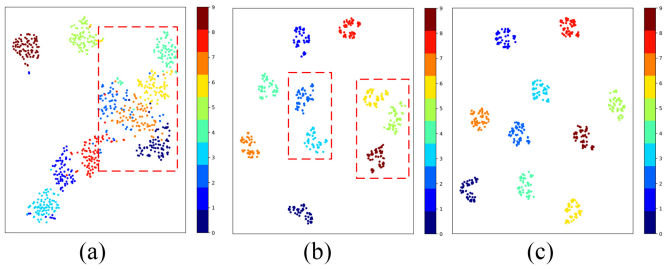


Fig. 14. Feature visualization of (a) radar-branch, (b) GEIs-branch, and (c) fusion scenario via t-SNE, respectively. Different identities are represented as different numbers with different colors. Here, we randomly choose ten identities from our test data set.

present whether to use the BPSA module, LSTRM module, single-scale fusion (SSF), and MSF module, respectively. For the single-modality branch, BPSA and LSTRM can effectively improve the performance under complex covariate conditions of the corresponding branch (from 78.151% to 80.264% and from 44.951% to 48.382% in terms of mean accuracy, respectively). For the multimodal fusion scenario, the baseline is the two-stream CNNs using feature segmentation with BA triplet loss and does not contain BPSA, LSTRM and MSF modules. The two-stream CNNs are only fused at the high-level feature space, i.e., SSF. Compared to the baseline model, we can see that all the proposed modules are designed reasonably and all modules can improve the performance. Furthermore, compared to a single modality branch, a single BPSA or LSTRM offers relatively little improvement in multimodal fusion scenario (+0.714% and +2.134% in terms of mean accuracy, respectively). The simultaneous introduction of BPSA and LSTRM has alleviated this dilemma (+2.844% in terms of mean accuracy). The results demonstrate that our BPSA and LSTRM can effectively mine complementary gait information of the two branches. As mentioned in the previous section, different layers of CNNs have different receptive fields and features of different levels focus on different information for gait recognition. Our MSF furthermore improves the gait recognition performance since multimodal and multiscale gait information are combined to obtain richer and more robust gait features. Combining all modules can achieve the best performance (+4.752% in terms of mean accuracy and +6.639% in terms of "CT" accuracy). In summary, the results of the ablation study demonstrate that our method can significantly improve the gait performance under cross-view and cross-walking-condition.

2) Visualization: In Fig. 14, we randomly choose ten identities from our test data set to visualize the feature distributions via t-SNE. Compared with single-modality branches, the feature distribution in the fusion scenario possesses better intraclass compactness and interclass differentiability. The identities are more distinguishable and easier to metric under the fusion scenario, which proves the more powerful feature representation ability of multimodal fusion.

V. CONCLUSION

In this article, a robust gait recognition approach with deep CNNs by fusing the camera and radar sensors is proposed for complex covariate conditions. An RGB camera and a millimeter-wave radar are used to collect GEIs and time-Doppler spectrograms. The GEIs portray a compact representation of the spatial body appearance over the gait cycle, while the time-Doppler spectrograms provide the

micro-motion pattern of gait. We propose a BPSA module to obtain more discriminative body part representations of GEIs. Meanwhile, we propose an LSTRM module to obtain the local and global micro-motion representation of time-Doppler spectrograms. Finally, we fuse the discriminative body part representation and the micro-motion pattern at the multiscale feature space to obtain richer and more robust gait features for human identification.

In the experiments, camera and radar data from 121 subjects with eight views and three walking conditions are collected. We further provide an extensive empirical evaluation in terms of various complex covariate conditions, namely, cross-view and cross-walking-condition. The experimental results demonstrate that camera and radar can capture highly complementary information for gait recognition, which can alleviate the different kinds of errors and the sensitivity to complex covariate conditions introduced by a single type of sensor. Besides that, we also explore the effect of different fusion ways on gait recognition performance. An interesting finding is that the appearance and micro-motion are highly complementary but completely two different information for gait recognition. Therefore, the fusion way that learning the similarities between these two types of gait information, such as Transformer, may be counterproductive.

In the future, we will continue to study gait recognition from two aspects. On the one hand, we will aim to improve the cross-view and cross-walking-condition accuracy of radar modality. On the other hand, we will study multimodal gait recognition under modality-absent conditions. When one modality does not work, such as cameras in dark or bad weather, how to make another modality mimic the features of the missing modality and achieve good fusion performance? Modality distillation may provide a good insight.

REFERENCES

[1] F. Deligianni, Y. Guo, and G. Yang, "From emotions to mood disorders: A survey on gait analysis methodology," *IEEE J. Biomed. Health Informat.*, vol. 23, no. 6, pp. 2302–2316, Nov. 2019.

[2] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.

[3] Z. Liu and S. Sarkar, "Improved gait recognition by gait dynamics normalization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 863–876, Jun. 2006.

[4] 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.

[5] Y. Zhang, G. Pan, K. Jia, M. Lu, Y. Wang, and Z. Wu, "Accelerometer-based gait recognition by sparse representation of signature points with clusters," *IEEE Trans. Cybern.*, vol. 45, no. 9, pp. 1864–1875, Sep. 2015.

[6] Q. Zou, L. Ni, Q. Wang, Q. Li, and S. Wang, "Robust gait recognition by integrating inertial and RGBD sensors," *IEEE Trans. Cybern.*, vol. 48, no. 4, pp. 1136–1150, Apr. 2018.

[7] A. Zhao et al., "Multimodal gait recognition for neurodegenerative diseases," *IEEE Trans. Cybern.*, vol. 52, no. 9, pp. 9439–9453, Sep. 2022.

[8] T. Connie, M. G. K. Ong, and A. B. J. Teoh, "A Grassmannian approach to address view change problem in gait recognition," *IEEE Trans. Cybern.*, vol. 47, no. 6, pp. 1395–1408, Jun. 2017.

[9] B. Vandersmissen et al., "Indoor person identification using a low-power FMCW radar," *IEEE Trans. Geosci. Remote. Sens.*, vol. 56, no. 7, pp. 3941–3952, Jul. 2018.

[10] V. Chen, "Radar micro-doppler signatures-principle and applications," *Radar Sci. Technol.*, vol. 10, no. 3, pp. 231–240, 2012.

[11] I. Rida, N. Al-Máadeed, and S. Al-Máadeed, "Robust gait recognition: A comprehensive survey," *IET Biom.*, vol. 8, no. 1, pp. 14–28, 2019.

[12] Z. Chen, G. Li, F. Fioranelli, and H. Griffiths, "Personnel recognition and gait classification based on multistatic micro-doppler signatures using deep convolutional neural networks," *IEEE Geosci. Remote. Sens. Lett.*, vol. 15, no. 5, pp. 669–673, May 2018.

[13] H. Li, A. Mehul, J. Le Kernec, S. Z. Gurbuz, and F. Fioranelli, "Sequential human gait classification with distributed radar sensor fusion," *IEEE Sensors J.*, vol. 21, no. 6, pp. 7590–7603, Mar. 2021.

[14] 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.

[15] C. Fan et al., "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14213–14221.

[16] X. Huang et al., "Context-sensitive temporal feature learning for gait recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 12889–12898.

[17] B. Lin, S. Zhang, and F. Bao, "Gait recognition with multiple-temporal-scale 3D convolutional neural network," in *Proc. ACM Multimedia*, 2020, pp. 3054–3062.

[18] 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.

[19] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017.

[20] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, "Gait recognition via semi-supervised disentangled representation learning to identity and Covariate features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13306–13316.

[21] K. Ren, L. Du, B. Wang, Q. Li, and J. Chen, "Statistical compressive sensing and feature extraction of time-frequency spectrum from Narrowband radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 56, no. 1, pp. 326–342, Feb. 2020.

[22] L. Du, L. Li, B. Wang, and J. Xiao, "Micro-doppler feature extraction based on time-frequency spectrogram for ground moving targets classification with low-resolution radar," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3756–3763, May 2016.

[23] L. Du, B. Wang, P. Wang, Y. Ma, and H. Liu, "Noise reduction method based on principal component analysis with beta process for micro-doppler radar signatures," *IEEE J. Sel. Topics Appl. Earth Observ. in Remote Sens.*, vol. 8, no. 8, pp. 4028–4040, Aug. 2015.

[24] Y. Li, L. Du, and H. Liu, "Hierarchical classification of moving vehicles based on empirical mode decomposition of micro-doppler signatures," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 5, pp. 3001–3013, May 2013.

[25] M. Z. Ozturk, C. Wu, B. Wang, and K. J. R. Liu, "GaitCube: Deep data cube learning for human recognition with Millimeter-wave radio," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 546–557, Jan. 2022.

[26] X. Bai, Y. Hui, L. Wang, and F. Zhou, "Radar-based human gait recognition using dual-channel deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 9767–9778, Dec. 2019.

[27] Y. Zhang, Y. Huang, S. Yu, and L. Wang, "Cross-view gait recognition by discriminative feature learning," *IEEE Trans. Image Process.*, vol. 29, pp. 1001–1015, 2020.

[28] S. Hou, C. Cao, X. Liu, and Y. Huang, "Gait lateral network: Learning discriminative and compact representations for gait recognition," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 382–398.

[29] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 14648–14656.

[30] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1602–1615, Jul. 2016.

[31] T. T. Verlekar, P. L. Correia, and L. D. Soares, "View-invariant gait recognition system using a gait energy image decomposition method," *IET Biometrics*, vol. 6, no. 4, pp. 299–306, 2017.

[32] P. Arora and S. Srivastava, "Gait recognition using gait Gaussian image," in *Proc. Int. Conf. Signal Process. Integr. Netw.*, 2015, pp. 791–794.

[33] A. Al-Tayyan, K. Assaleh, and T. Shanableh, "Decision-level fusion for single-view gait recognition with various carrying and clothing conditions," *Image Vis. Comput.*, vol. 61, pp. 54–69, May 2017.

[34] R. Atta, S. Shaheen, and M. Ghanbari, "Human identification based on temporal lifting using 5/3 wavelet filters and radon transform," *Pattern Recognit.*, vol. 69, pp. 213–224, Sep. 2017.

[35] Y. Kim and H. Ling, "Human activity classification based on micro-doppler signatures using a support vector machine," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 5, pp. 1328–1337, May 2009.
[36] C. Karabacak, S. Z. Gurbuz, A. C. Gürbüz, M. B. Guldogan, G. Hendeby, and F. Gustafsson, "Knowledge exploitation for human micro-doppler classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 10, pp. 2125–2129, Oct. 2015.
[37] F. Fioranelli, M. Ritchie, S. Z. Gürbüz, and H. D. Griffiths, "Feature diversity for Optimized human micro-doppler classification using Multistatic radar," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 53, no. 2, pp. 640–654, Apr. 2017.
[38] P. Cao, W. Xia, M. Ye, J. Zhang, and J. Zhou, "Radar-ID: Human identification based on radar micro-doppler signatures using deep convolutional neural networks," *IET Radar Sonar Navigat.*, vol. 12, no. 7, pp. 729–734, 2018.
[39] Y. Lang, Q. Wang, Y. Yang, C. Hou, H. Liu, and Y. He, "Joint motion classification and person identification via Multitask learning for smart homes," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 9596–9605, Dec. 2019.
[40] Y. Yang, C. Hou, Y. Lang, G. Yue, Y. He, and W. Xiang, "Person identification using micro-doppler signatures of human motions and UWB radar," *IEEE Microw. Wireless Compon. Lett.*, vol. 29, no. 5, pp. 366–368, May 2019.
[41] P. Addabbo, M. L. Bernardi, F. Biondi, M. Cimitile, C. Clemente, and D. Orlando, "Gait recognition using FMCW radar and temporal convolutional deep neural networks," in *Proc. IEEE Int. Workshop Metrol. Aerosp.*, 2020, pp. 171–175.
[42] Y. Cheng and Y. Liu, "Person reidentification based on automotive radar point clouds," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, May 2021, Art. no. 5101913.
[43] L. Zhang, C. Wang, M. Ma, and D. Zhang, "WiDIGR: Direction-independent gait recognition system using commercial Wi-Fi devices," *IEEE Internet Things J.*, vol. 7, no. 2, pp. 1178–1191, Feb. 2020.
[44] Y. Cao, Z. Zhou, C. Zhu, P. Duan, X. Chen, and J. Li, "A lightweight deep learning algorithm for WiFi-based identity recognition," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17449–17459, Dec. 2021.
[45] D. Wang, J. Yang, W. Cui, L. Xie, and S. Sun, "CAUTION: A robust WiFi-based human authentication system via few-shot open-set recognition," *IEEE Internet Things J.*, vol. 9, no. 18, pp. 17323–17333, Sep. 2022.
[46] F. Sun, C. Mao, X. Fan, and Y. Li, "Accelerometer-based speed-adaptive gait authentication method for wearable IoT devices," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 820–830, Feb. 2019.
[47] P. Kumar, S. Mukherjee, R. Saini, P. Kaushik, P. P. Roy, and D. P. Dogra, "Multimodal gait recognition with inertial sensor data and video using evolutionary algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 27, no. 5, pp. 956–965, May 2019.
[48] L. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with Atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.
[49] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 1538–1543.
[50] P. Zhang, L. Hu, B. Zhang, P. Pan, and D. Alibaba, "Spatial consistent memory network for semi-supervised video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, vol. 6, 2020, p. 2.
[51] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, "Learning discriminative features with multiple Granularities for person re-identification," in *Proc. ACM Int. Conf. Multimedia*, 2018, pp. 274–282.
[52] A. Hermans, L. Beyer, and B. Leibe. "In defense of the triplet loss for person re-identification," 2017. [Online]. Available: http://arxiv.org/abs/1703.07737.
[53] Z. Xia, G. Ding, H. Wang, and F. Xu, "Person identification with Millimeter-wave radar in realistic smart home scenarios," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, Oct. 2021, Art. no. 3509405.
[54] H. G. Doherty, R. A. Burgueño, R. P. Trommel, V. Papanastasiou, and R. I. Harmanny, "Attention-based deep learning networks for identification of human gait using radar micro-doppler spectrograms," *Int. J. Microw. Wireless Technol.*, vol. 13, no. 7, pp. 734–739, 2021.
[55] A. Vaswani et al., "Attention is all you need," in *Proc. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
[56] Y. Kim, J. W. Choi, and D. Kum, "GRIF Net: Gated region of interest fusion network for robust 3D object detection from radar point cloud and monocular image," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2020, pp. 10857–10864.
[57] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "TransFuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Aug. 19, 2022, doi: 10.1109/TPAMI.2022.3200245.
[58] X. Hao, S. Zhao, M. Ye, and J. Shen, "Cross-modality person re-identification via modality confusion and center aggregation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 16383–16392.

Yu Shi (Graduate Student Member, IEEE) received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2019, where he is currently pursuing the Ph.D. degree in signal processing with the National Laboratory of Radar Signal Processing.

His research interests include radar automatic target recognition and deep cross-domain learning.

Lan Du (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in July 2001, March 2004, and June 2007, respectively.

Her doctoral dissertation was granted National Excellent Doctoral Dissertation of PR China in 2009. From September 2007 to September 2009, she did research work as a Postdoctoral Research Associate with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. She is currently a Professor with the National Laboratory of Radar Signal Processing, Xidian University. Her research work is supported by NSFC for Excellent Young Scholars. Her main research interests are in the fields of statistical signal processing and machine learning with application to radar target recognition.

Xiaoyang Chen received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2020, where he is currently pursuing the M.S. degree in signal processing with the National Laboratory of Radar Signal Processing.

His research interests include radar signal processing and gait recognition.

Xun Liao received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2019, where he is currently pursuing the M.S. degree in electronic information with the National Laboratory of Radar Signal Processing.

His research interests include radar automatic target recognition and deep learning.

Zengyu Yu received the B.Eng. degree in electronic and information engineering from Beijing Institute of Technology, Beijing, China, in 2018, and the M.S. degree in signal processing with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China, in 2021.

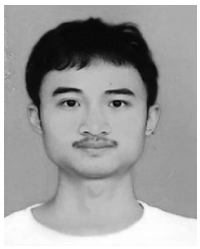
His research interests include radar signal processing and radar target recognition.

Zenghui Li received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2019, and the M.S. degree in signal processing from the National Laboratory of Radar Signal Processing, Xidian University, in 2022.

His research interests include radar signal processing and radar target recognition.

Chunxin Wang received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2019, and the M.S. degree in signal processing from the National Laboratory of Radar Signal Processing, Xidian University, in 2022.

His research interests include radar signal processing and radar target track.



Shikun Xue received the B.Eng. degree in electronic and information engineering from Xidian University, Xi'an, China, in 2021, where he is currently pursuing the M.S. degree in signal processing with the National Laboratory of Radar Signal Processing.

His research interests include radar gait recognition and radar action recognition.