# RPNet: Gait Recognition With Relationships Between Each Body-Parts

Hao Qin, Zhenxue Chen, Qingqiang Guo, Q. M. Jonathan Wu, *Senior Member, IEEE*, and Mengxu Lu

***Abstract*— At present, many studies have shown that partitioning the gait sequence and its feature map can improve the accuracy of gait recognition. However, most models just cut the feature map at a fixed single scale, which loses the dependence between various parts. So, our paper proposes a structure called Part Feature Relationship Extractor (PFRE) to discover all of the relationships between each parts for gait recognition. The paper uses PFRE and a Convolutional Neural Network (CNN) to form the RPNet. PFRE is divided into two parts. One part that we call the Total-Partial Feature Extractor (TPFE) is used to extract the features of different scale blocks, and the other part, called the Adjacent Feature Relation Extractor (AFRE), is used to find the relationships between each block. At the same time, the paper adjusts the number of input frames during training to perform quantitative experiments and finds the rule between the number of input frames and the performance of the model. Our model is tested on three public gait datasets, CASIA-B, OU-LP and OU-MVLP. It exhibits a significant level of robustness to occlusion situations, and achieves accuracies of 92.82% and 80.26% on CASIA-B under BG # and CL # conditions, respectively. The results show that our method reaches the top level among state-of-the-art methods.**

***Index Terms*— Gait recognition, convolutional neural network (CNN), partial relationship, different scale blocks.**

Manuscript received March 16, 2021; revised June 8, 2021 and June 20, 2021; accepted June 30, 2021. Date of publication July 7, 2021; date of current version May 5, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61876099, in part by the National Key Research and Development Program of China under Grant 2019YFB1311001, in part by the Scientific and Technological Development Project of Shandong Province under Grant 2019GSF111002, in part by the Shandong Provincial Key Research and Development Program through the Major Scientific and Technological Innovation Project under Grant 2019JZZY010119, in part by the Foundation of the Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education under Grant 2018ICIP03, and in part by the Foundation of the State Key Laboratory of Integrated Services Networks under Grant ISN20-06. This article was recommended by Associate Editor F. M. Zhu. *(Hao Qin and Zhenxue Chen contributed equally to this work.) (Corresponding author: Zhenxue Chen.)*

Hao Qin, Qingqiang Guo, and Mengxu Lu are with the School of Control Science and Engineering, Shandong University, Jinan 250061, China (e-mail: 202014785@mail.sdu.edu.cn; gqq@sdu.edu.cn; 201500171046@mail.sdu.edu.cn).

Zhenxue Chen is with the School of Control Science and Engineering, Shandong University, Jinan 250061, China, and also with the State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China (e-mail: chenzhenxue@sdu.edu.cn).

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3095290.

Digital Object Identifier 10.1109/TCSVT.2021.3095290

## I. Introduction

NOWADAYS, biometric technology plays an important role in personal identification, but it also has many problems. For example, face recognition is easily affected by objective factors such as makeup, age, and distance, and fingerprints are easy to forge. Gait recognition is a kind of biometric identification technology that can judge individual identity by walking posture [1], [2]. Gait recognition can be done from a long distance and does not require high-resolution input images. It is imperceptible and non-invasive, and has large development potential in the field of security, monitoring and criminal investigation [3], [4].

However, gait recognition is easily affected by occlusion, different viewing angles and appearance. Scholars have done a lot of research in view of these problems. Bashir *et al.* [5] used a novel gait representation termed Gait Entropy Image (GEI). It mainly captures motion information and is robust to changes in covariable conditions that affect appearance. Ben *et al.* [6] aligned GEI across views with the coupled bilinear discriminant projection (CBDP), which solves the problem of not being able to transfer cross-view gait features into similar ones.

At the same time, in order to effectively match gait pairs across different views, they proposed a coupled patch alignment (CPA) algorithm [7]. And they proposed a general tensor representation framework to recognize cross-view gait [8]. Aggarwal and Vishwakarma [9] used AESI, Zernike moment invariants (ZMIs), SDOGs and MDPs to detect occlusion covariates and reduce their effects, thereby improving the accuracy of gait recognition. Wu *et al.* [10] studied a gait-based pedestrian recognition method based on deep convolutional neural networks (CNNs) similarity learning, becoming the first study on gait recognition based on deep CNN. Chao *et al.* [11] proposed a new network named GaitSet to identify information from the set of independent frames. This greatly improves the flexibility of gait recognition. To remedy confounding variables, Zhang *et al.* [12] proposed a new auto-encoder framework to explicitly separate posture and appearance features from RGB images; over time, LSTM-based posture features integrate to generate gait features. Wolf *et al.* [13] proposed a deep convolutional neural network that uses 3D convolution to capture spatio-temporal features in multiple views. Fan *et al.* [14] fully considered the influence of different body parts of pedestrians on gait recognition and proposed a new network named GaitPart. It greatly improves the accuracy of gait recognition. Xu *et al.* [15]

1051-8215 © 2021 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

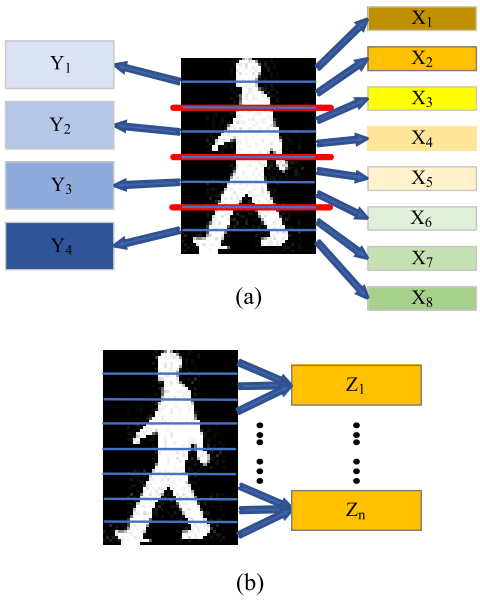



Fig. 1. **(a)**: The human body is divided into different scales (4 & 8), and different characteristics can be obtained. **(b)**: After combining the adjacent parts of the human body, a new feature can be obtained, and used for personal identification.

proposed a pairwise spatial transformation network (PSTN) for cross-view gait recognition, which can reduce unnecessary feature misalignment due to viewpoint differences before the recognition step.

At the same time, the development of gait recognition is greatly influenced by other fields, such as the field of Human Action Recognition (HAR) and Person Re-Identification (ReID). Some work in the field of HAR [16]–[19] has enriched the methods of analyzing human posture in gait recognition. And some work in the field of ReID [20]–[24] has enriched the methods of image clustering in gait recognition where the intra-class spacing is greater than the inter-class spacing.

With the development of gait recognition research, accuracy has reached a very high level under normal conditions (96%) [14]. However, the synergy between various parts of the human body when walking is always ignored. Each part of the human body can be individually identified, and the coordination relationship between each part of each person is different (see Fig. 1). This relationship can also be used as a condition for identification. When there is a large occlusion in the view, the recognition accuracy will decrease rapidly. The common use of multiple discrimination conditions may reduce the loss of conditions caused by occlusion. In order to improve the recognition effect when there is occlusion, we propose a model named RPNet. RPNet fully considers the relationship between different scale parts of the human body, and then integrates them by simple splicing. It improves the recognition accuracy under the condition of occlusion and reaches the leading level.

Compared with the previous work, RPNet has three main novelties in design: (1) RPNet compresses feature maps at different scales to obtain different fine-grained feature blocks. (2) RPNet uses a concise convolutional layer and pooling layer to extract the relationship features between adjacent parts of the feature map. (3) RPNet uses an encoder with fewer convolutional layers.

As shown in Fig. 2, RPNet first extracts several partial features of gait profile by a convolutional network, and part of the features are then aggregated in two separate ways. One branch (TPFE) is responsible for mapping feature vectors of different sizes. The other branch (AFRE) is responsible for finding the synergy between adjacent parts. Then, the features of these two branches are stitched together to obtain the complete output.

All in all, the contributions of our work are mainly as follows:

- We propose a module that maps feature vectors to a more recognizable space, called PFRE, in order to find the relationships between each part of the feature vector.
- We do a quantitative experiment and analyze the experimental results to improve the interpretability of gait recognition in deep learning.
- We reorganize the structure of the OULP and OUMVLP datasets to make them easy to train with the CASIA-B dataset.
- Our model performs well on **CASIA-B** [25], **OULP** [26] and **OUMVLP** [27] datasets, showing high robustness to occlusion conditions and generalization ability to large amounts of data.

## II. Related Works

In this section, we will briefly introduce gait recognition and partial feature processing methods.

### A. Gait Recognition

According to the selected types of gait features, gait recognition methods can be roughly divided into model-based categories and sequence-based categories. The reason for this classification is that the gait data itself contains two components: the structured component and the dynamic component.

The model-based approach is to model the human body or movement through structured components during walking. According to the extracted feature dimension, the method is subdivided into a two-dimensional motion model and a three-dimensional motion model. The most obvious feature of walking is the movement of the legs. Many two-dimensional models can achieve good results only on the leg modeling. Jean *et al.* [28] also used the movements of the head and feet as gait data. When walking, the feet will bring the occlusion problem, so the moment of occlusion is used as the alternating point of left and right feet.

The two-dimensional model will expose its own defects in the case of multiple perspectives or shielding and cannot fully express the gait. Therefore, the three-dimensional human body structure model is needed to solve this problem. Wei Lu *et al.* [29] proposed a gait recognition method based on joint distribution of motion angles. They used a number of key nodes to describe the motion posture, connecting the corresponding angles of the left and right legs to obtain the distribution spectrum of the joint and establishing the

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

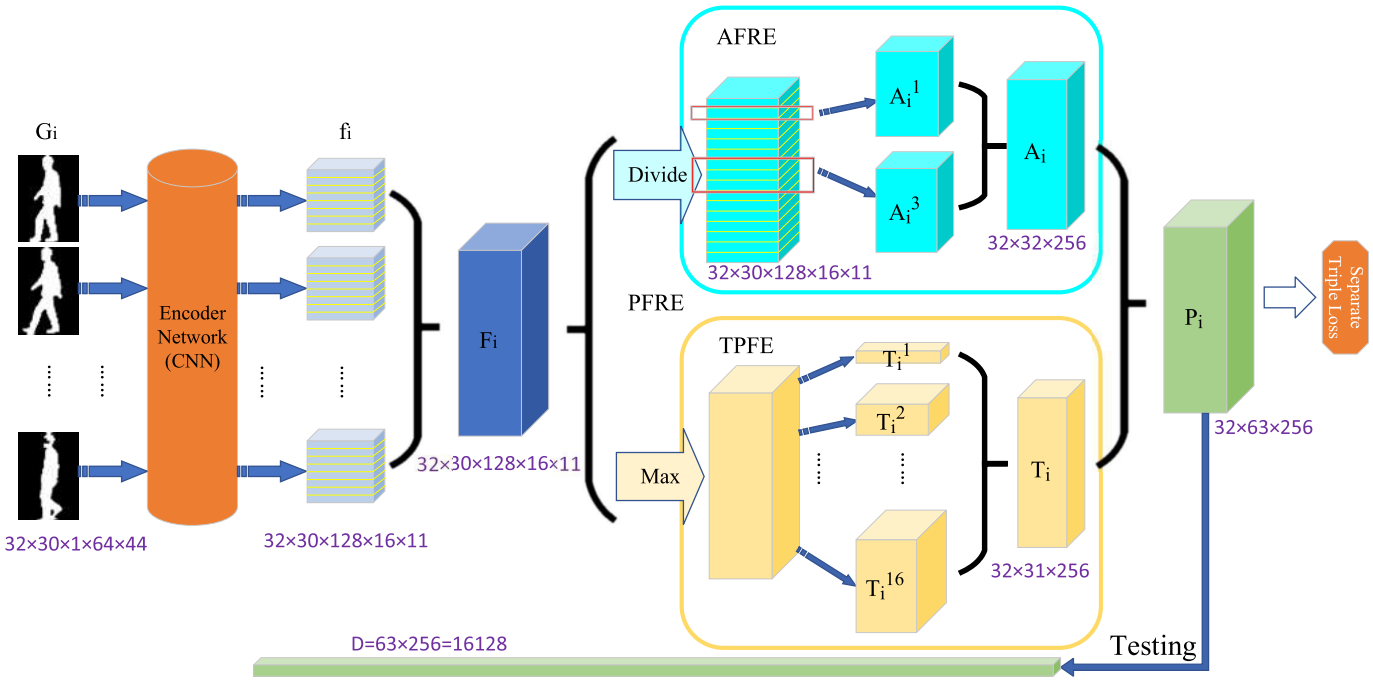



Fig. 2. **The framework of RPNet.** Our framework consists of two parts. **Part A** is a feature extraction network composed of a convolutional neural network, represented by an orange cylinder in the figure. Different pictures share a set of parameters in the encoder network. **Part B** consists of AFRE and TPFE; we call it PFRE. Part B consists of the convolutional layer, the maximum pooling layer, the average pooling layer, and dimensional operations. '$G_i$' represents the input sequence. '$P_i$' represents the final output feature. When testing, the feature vector of each individual in $P_i$ is expanded into one dimension. The purple numbers represent the dimensions of the vector.

feature histogram based on the joint distribution of angles. Deng and Wang [30] proposed a new model-based gait recognition method by employing deterministic learning theory to capture the gait dynamics underlying Kinect-based gait parameters. In the future, the development of model-based approaches may depend on advances in human pose estimation algorithms [31], [32].

Sequence-based gait recognition obtains gait data by characterizing the whole movement pattern of the human body; this method is divided into gait energy image method and gait image sequence method according to different research objects. Ju and Bhanu [33] proposed the Gait Energy Image (GEI) to characterize the human gait. GEI realizes the representation of human motion sequence in a single image while preserving time information. GEI plays a very important role in gait recognition, and many researchers have made improvements based on it [34]–[37]. Considering the diversity of gait images, Zhou *et al.* [38] proposed a kernel-based semantic hashing (KSH) model to increase the weight of images with similar styles by optimizing a semantic triplet ranking loss.

In a periodic gait sequence, the gait at each moment is different, and the continuous gait pose at each moment contains the gait information of each person. The spatial information of each part of the human body in the sequence can be analyzed by representing the gait data with a set of binary profile images in a continuous period of one week, and then the identity can be recognized [10], [13], [14], [39]. The method based on sequence does not need to establish a prior model, but only carries on the statistical analysis of the spatial and temporal characteristics of the gait image, which has a low computational complexity. However, the problem of occlusion may occur when the image is acquired in real life.

### *B. Partial Feature Processing*

'Partial feature' refers to the small piece of feature vector obtained after cutting the general feature vector. Partial feature contains more fine-grained information, which improves the identification of feature vectors.

In the field of person re-identification, there has been much research on the block processing of image or feature map [20]–[23], [40]–[42]. Chao *et al.* [11] applied HPP to gait recognition and cut the feature maps extracted from the convolutional network. Fan *et al.* [14] simply cut the feature maps of the convolution process into small pieces and convolved each piece separately. The methods they proposed have greatly improved the accuracy of gait recognition.

Blocking the feature vector allows for more fine-grained features, but it also loses the dependencies between each parts. Based on the above research, we design a structure called Part Feature Relationship Extractor (**PFRE**) to deal with the characteristics of each part. PFRE fully considers the relationship between the whole and each parts of the feature graph and the relationships between the close parts. PFRE automatically adjusts the weight of each relationship through the neural network.

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

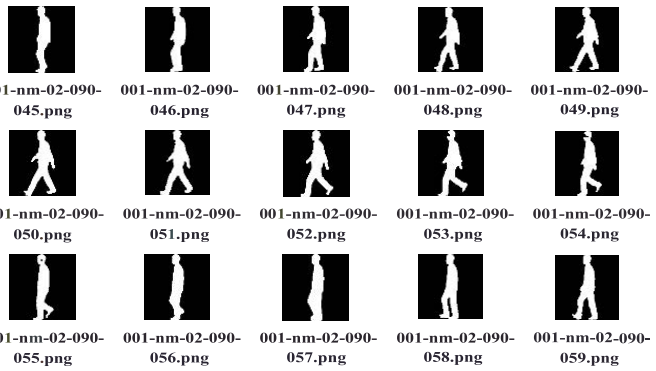



Fig. 3. A set of gait silhouettes from **CASIA-B**.

## III. METHOD

In this section, we detail the model we proposed and the collaboration between its parts. The overall pipeline is illustrated in Fig. 2.

### A. Problem Formulation

The input of RPNet is a set of gait silhouettes $G_i$ (see Fig. 3), and it has a dimension of $P*K*C*H*W$, where, $P$ represents the number of the sets input, $K$ represents the number of frames of the image in each pedestrian sequence, $C$ represents the number of channels for the input image, and $H$, $W$ represents its height and width.

Let a gait sequence be $G_i = \{g_i | i = 1, 2, \ldots, n\}$, where $g_i$ represents a frame of gait image, and $n$ represents the number of input frames. To get the feature map $F_i$, we input $G_i$ into a convolutional neural network $F$:

$$F_i = F(G_i), \tag{1}$$

where, $F$ contains several convolutional layers and pooling layers, and focal convolution is done for the intermediate feature graph. $F_i$ is still a five-dimensional vector, and the meaning of each dimension remains the same.

In order to extract the local information in $F_i$ and improve the identification of feature vector, we map $F_i$ with PFRE:

$$P_i = P(F_i), \tag{2}$$

where $P$ stands for PFRE. Actually, PFRE do two parallel manipulations for $F_i$:

$$T_i = T(F_i) \tag{3}$$

$$A_i = A(F_i), \tag{4}$$

where $T$ represents the Total-Partial Feature Extractor (**TPFE**) and $A$ represents the Adjacent Feature Relation Extractor (**AFRE**). Lastly, $T_i$ is spliced with $A_i$, and we get the final characteristic graph $P_i$:

$$P_i = \{T_i, A_i\}, \tag{5}$$

where '{}' means splicing operation.

The dimensions of $T_i$ and $A_i$ are $P*M_1*C$ and $P*M_2*C$, respectively, where $P$ and $C$ still represent the number of people and channels. $M_1$ and $M_2$ are hyperparameters that represent how many feature blocks are stacked in the feature vector that is output by TPFE and AFRE, respectively.

TABLE I
THE OPERATION AND ITS PARAMETERS OF CONVOLUTION AND POOLING

| size | operation | parameter |
|---|---|---|
| 3 | Conv | kernel_size=3, padding=1, stride=1 & kernel_size=1, padding=0, stride=1 |
| | AvgPool | kernel_size=3, padding=1, stride=1 |
| | MaxPool | |
| 5 | Conv | kernel_size=3, padding=1, stride=1 & kernel_size=3, padding=1, stride=1 |
| | AvgPool | kernel_size=5, padding=2, stride=1 |
| | MaxPool | |

### B. Adjacent Feature Relation Extractor

The Adjacent Feature Relation Extractor (**AFRE**), aiming to extract fine-grained features from feature vectors without losing the dependence between adjacent feature blocks, is composed of two little branches. The first branch uses a box of scale 1 to traverse the feature vector, and the second branch uses a box of scale 3. Next, AFRE will be described in detail (see Fig. 4).

*1) Definition:* AFRE is a feature extractor that contains multiple parallel convolution operations. No parameters are shared among the parallel convolution kernels.

*2) Motivation:* The various parts of the human body are connected, and there is a dependent relationship between each part, especially between adjacent parts. If we are only convolving each piece, we will lose that dependency. So, we aggregate adjacent blocks and convolve them.

*3) Operation:* Firstly, we take the sum of the maximum value and the average value in the fourth dimension of $F_i$ to obtain $\tilde{F}_i$ and then divide the feature vector $\tilde{F}_i$ into $m$ small pieces to get $f_i$. After that, we do convolution operation and pooling operation for each small feature block separately:

$$\tilde{a}_i^1 = conv(f_i), \tag{6}$$

$$\bar{a}_i^1 = MP(f_i) + AP(f_i). \tag{7}$$

The parameters in (6) are not shared. Then, multiply $\tilde{a}_i^1$ and $\bar{a}_i^1$:

$$a_i'^1 = \tilde{a}_i^1 \times \bar{a}_i^1, \tag{8}$$

when performing convolution operations and pooling operations, we use cores of size 3 and 5, respectively. In other words, the operations in the blue box in Fig. 4 will be done twice so that we can get $a_i'^1$ and $a_i''^1$. The specific operation and its parameters are shown in Tab. I. After summing $a_i'^1$ and $a_i''^1$, take the maximum value in the second dimension to obtain $a_i^1$:

$$a_i^1 = max(a_i'^1 + a_i''^1). \tag{9}$$

All $a_i^1$ is spliced to get $A_i^1$. We merge the operations in (6)(7)(8)(9) into $A^1$, which is:

$$A_i^1 = A^1(\tilde{F}_i). \tag{10}$$

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

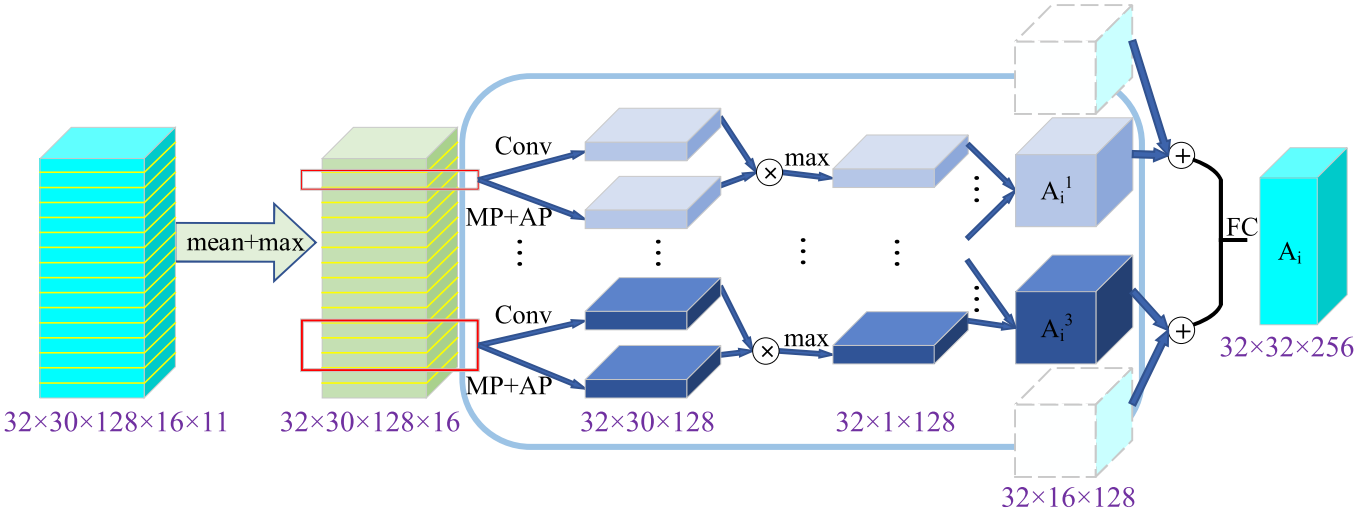



Fig. 4. The detailed structure of Adjacent Feature Relation Extractor (**AFRE**). 'Conv' represents convolution operation. Parallel convolution operations do not share parameters. 'MP' and 'AP' represent maximum pooling and average pooling, respectively. 'FC' represents a fully connected layer. The purple numbers represent the dimensions of the vector.

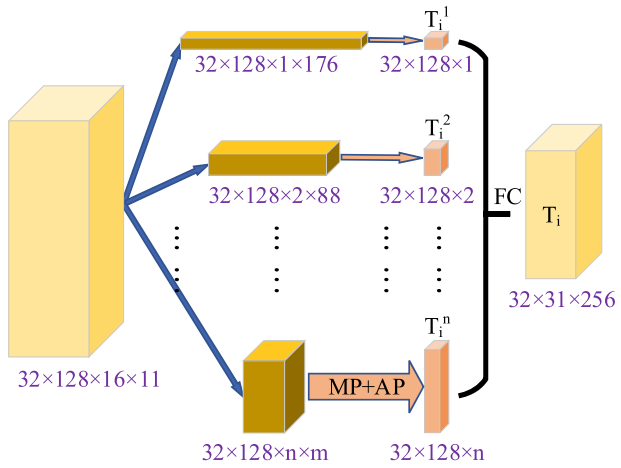



Fig. 5. The detailed structure of Total-Partial Feature Extractor (**TPFE**). 'MP' and 'AP' represent maximum pooling and average pooling, respectively. 'FC' represents a fully connected layer. The purple numbers represent the dimensions of the vector.

At the same time, we aggregate the three adjacent small pieces in $\tilde{F}_i$ and then do the same operations as $A^1$ on them:

$$A_i^3 = A^3(\tilde{F}_i). \tag{11}$$

The method of aggregation is direct addition. The characteristic blocks at both ends are multiplied by two and added to an adjacent block to keep the data stable.

Finally, $A_i^1$ and $A_i^3$ are spliced to get $A_i^{\{1,3\}}$. We obtain $A_i$ after going through a full connection layer:

$$A_i = FC(\{A_i^1, A_i^3\}). \tag{12}$$

### C. Total-Partial Feature Extractor

As shown in Fig. 5, the Total-Partial Feature Extractor (**TPFE**) contains 5 branches, and different branches perform averaging and maximizing operations at different scales. Next, the TPFE will be described in detail.

*1) Definition:* TPFE is a feature extractor that includes multiple maximum-value and average-value modules of different scales.

*2) Motivation:* Walking is a common behavior in daily life, but it contains complex movement mechanisms. To find all of the parts that might contribute to the identification of a pedestrian, we use TPFE to process the feature graph.

*3) Operation:* As shown in Fig. 2, we perform a maximum operation on $F_i$ to obtained $\tilde{F}_i'$ before sending it to TPFE, which reduces its dimension. TPFE will do five different maximum and average operations for $\tilde{F}_i'$. Firstly, TPFE compresses the $H$ dimension to $n$ and expands the $W$ dimension to $H*W/n$:

$$t_i^n = t^n(\tilde{F}_i^1). \tag{13}$$

Then, TPFE takes the maximum and average values of the $H*W/n$ dimension:

$$T_i^n = max(t_i^n) + ave(t_i^n). \tag{14}$$

We repeat these two parts five times with different $n$. Finally, all of the $T_i^n$ are going to be spliced together. Through a full connection layer, we will get $T_i$:

$$T_i = FC(\sum_n T_i^n). \tag{15}$$

### D. Training and Testing

*1) Training:* The final feature vector of the whole network is $P_i$ with three dimensions. The corresponding characteristics between different samples are used to calculate the loss. In this paper, we use the Batch All (BA+) triple loss function [43] to train the network. At each iteration, RPNet selects $I$ principals and $J$ sequences from each principal. The Euclidean distance between the feature vectors of the same subject is defined as

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

*PD*, and the Euclidean distance between the feature vectors of different subjects is defined as *ND*. The loss is going to be:

$$Loss = max(PD - ND + a, 0) \tag{16}$$

where $a$ is a hyperparameter.

*2) Testing:* During the test session, Gallery and Probe are fed into RPNet. For the set of feature vectors obtained, the two feature vectors closest to each other in Euclidean space are considered as origination from the same subject. The accuracy is calculated from this.

## IV. Experiments

In this section, we will introduce the dataset used to train the model, the selection of various parameters during model training, and some experimental results.

### A. Datasets and Training Details

*1) CASIA-B:* Contains 124 sequences recorded by different subjects, CASIA-B is a widely used gait dataset. Each sequence can be divided into 10 categories based on what is worn by the pedestrians. These include: NM #01-06, BG #01-02, and CL #01-02. Where NM stands for 'normal condition', BG stands for 'backpack condition', and CL stands for 'coat condition'. Each category is then divided into 11 smaller sequences according to the angle from 0° to 180°. When training, we use the first 74 subjects. In testing, we use the last 50 subjects. In the test, the first four groups in normal condition (NM #01-04) are used as galleries. The remaining NM #05-06, BG #01-02, and CL #01-02 are used as probes in three cases.

*2) OULP:* The OULP dataset consisted of 4,007 subjects, including 2,135 men and 1,872 women, ranging in age from 1 to 94 years. Eight sequences of four angles (55°, 65°, 75°, 85°) are performed for each subject and two sequences for each angle. One serves as a gallery and the other as a probe. OULP contains fewer variables but more principals than CASIA-B. Therefore, the OULP dataset is a test of the generalizability of the model.

*3) OUMVLP:* The OUMVLP dataset consisted of 10,307 subjects, including 5,114 men and 5,193 women, ranging in age from 2 to 87 years. Twenty-eight sequences of fourteen angles (0°, 15°, …, 90°, 180°, 195°, …, 270°) are performed for each subject and two sequences for each angle. One serves as a gallery and the other as a probe.

*4) Training Details:* For the CASIA-B and OUMVLP datasets, the method mentioned in [11] is used for preprocessing, and the image size obtained after processing is 66 × 44. For the OULP dataset, we changed the size of the image from 128 × 88 to 64 × 44 directly by scaling. For the convenience of training, we change the structure of the OULP and OUMVLP datasets as follows: Pid/Seq/Angle, where Pid is the label of the subject. Seq is divided into Seq00 (gallery) and Seq01 (probe). Angle contains 55°, 65°, 75°, 85° and 0°, 15°, …, 90°, 180°, 195°, …, 270°, representing the angle of the subject. The optimizer selects Adam [44], whose learning rate is set to 1e-4 (1e-5 when the OUMVLP dataset is trained 150K iterations) and momentum is set to 0.9. The margin in the BA+ triple loss is set to 0.2. The parameter $n$ in TPFE is set to [1, 2, 4, 8, 16].

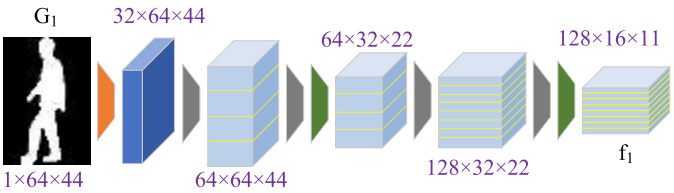



Fig. 6. The specific structure of our encoder network. For ease of presentation, the figure only shows the processing of one frame of image. The orange trapezoid represents the Convolution. The gray trapezoid represents the **Focal Convolution**. The green trapezoid represents pooling. The purple numbers represent the dimensions of the vector.

In CASIA-B, we randomly select 4 subjects for training at a time and 8 sequences from each subject. For the convenience of comparison, the data in Tab. II is obtained by selecting 30 frames from each sequence for training. Interestingly, we find that the change in the number of selected frames during training would regularly affect the performance of RPNet. With the increase in the number of selected frames, the accuracy of strong occlusion will be improved. Detailed data and analysis will be presented in the experimental section. In OULP and OUMVLP, we randomly select 16 subjects for training at a time and 8 sequences from each subject. The other parameters are the same as those in the CASIA-B dataset. Due to the powerful generalization performance of RPNet, its accuracy rate on the OULP dataset has reached 99.8%, so we do not do too many experiments on this dataset. All of the models are trained with 2 NVIDIA 2080TI GPUS. We trained our model 180K iterations on the CASIA-B dataset, 40K iterations on the OULP dataset, and 250K iterations on the OUMVLP dataset.

*5) Encoder Network Structure:* Our feature encoder network consists of ordinary convolution, pooling, and block convolution. The specific structure is shown in Fig. 6.

The encoder network that we use is similar to [11], [14], but slightly adjusted. Focal Convolution is proposed by [14]. Its specific operation is to cut the input feature vector during convolution and convolve each part separately. The core idea of Focal Convolution is to make the advanced convolution kernel focus on more local details in each specific part of the input frame, thereby intuitively using more fine-grained local information. When training on the OUMVLP dataset, a Focal Convolution layer is added at the end of the encoder to make the number of output channels change to 256, and the remaining parameters remain unchanged.

### B. Comparison With the State-of-the-Art

*1) CASIA-B:* Tab. II shows the experimental results of other state-of-the-art models and RPNet. The accuracies of all compared models are extracted from their papers. In order to make the results more convincing, we show the accuracy of 11 angles under three conditions. The accuracies are averaged on the gallery view, so the variable is only the left probe view.

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

TABLE II
AVERAGED RANK-1 ACCURACIES ON **CASIA-B** WITH CROSS CONDITIONS

| Gallery NM #1-4 | | 0°-180° | | | | | | | | | | | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Probe | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | |
| NM #5-6 | GaitNet[12] | 91.2 | - | - | 95.6 | - | 92.6 | - | 96.0 | - | - | - | 93.9 |
| | CNN-Ensemble[10] | 88.7 | 95.1 | 98.2 | 96.4 | 94.1 | 91.5 | 93.9 | 97.5 | 98.4 | 95.8 | 85.6 | 94.1 |
| | GaitSet[11] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
| | GaitPart[14] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | **99.2** | 97.8 | 90.4 | 96.2 |
| | MvGGAN[45] | 94.8 | **99.0** | **99.7** | **99.2** | **96.6** | **93.7** | **96.3** | **98.6** | **99.2** | **98.2** | **92.3** | **97.1** |
| | RPNet(Ours) | **95.1** | **99.0** | 99.1 | 98.3 | 95.7 | 93.6 | 95.9 | 98.3 | 98.6 | 97.7 | 90.8 | 96.55 |
| BG #1-2 | GaitNet[12] | 83.0 | - | - | 86.6 | - | 74.8 | - | 85.8 | - | - | - | 82.6 |
| | CNN-LB[10] | 64.2 | 80.6 | 82.7 | 76.9 | 64.8 | 63.1 | 68.0 | 76.9 | 82.2 | 75.4 | 61.3 | 72.4 |
| | GaitSet[11] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
| | GaitPart[14] | 89.1 | 94.8 | 96.7 | **95.1** | 88.3 | 84.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 91.5 |
| | MvGGAN[45] | **92.4** | 94.7 | **97.2** | 94.6 | 88.7 | 83.6 | 87.8 | 93.8 | **96.3** | **95.2** | **86.8** | 91.9 |
| | RPNet(Ours) | 92.3 | **96.6** | 96.6 | 94.5 | **91.9** | **87.6** | **90.7** | **94.7** | 96.0 | 93.9 | 86.1 | **92.82** |
| CL #1-2 | GaitNet[12] | 42.1 | - | - | 70.7 | - | 70.6 | - | 69.4 | - | - | - | 63.2 |
| | CNN-LB[10] | 37.7 | 57.2 | 66.6 | 61.1 | 55.2 | 54.6 | 55.2 | 59.1 | 58.9 | 48.8 | 39.4 | 54.0 |
| | GaitSet[11] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.3 | 68.4 | 50.0 | 70.4 |
| | GaitPart[14] | 70.7 | 85.5 | 86.9 | **83.3** | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | **80.2** | 66.5 | 78.7 |
| | MvGGAN[45] | 70.5 | 77.9 | 82.5 | 82.7 | 77.4 | 73.6 | 73.8 | 77.8 | 77.6 | 72.5 | 64.8 | 75.6 |
| | RPNet(Ours) | **75.6** | **87.1** | **88.3** | 83.1 | **78.8** | **78.0** | **79.9** | **82.7** | **83.9** | 78.9 | **66.6** | **80.26** |

The results in Tab. II show that our model achieves the highest accuracy under BG # and CL # conditions. The original intention of our method is to enhance the learning ability under occlusion. And the experimental results show that the accuracy of BG # and CL # is improved, demonstrating the effectiveness of our method. The accuracy rate in the NM # condition does not have an absolute advantage, but it does not have much disadvantage compared with other top models. MvGGAN first expands the dataset through the GAN network, and then uses the end-to-end network for clustering. When the dataset is small and uncomplicated, this method can greatly improve the robustness of the model and avoid overfitting. However, our method only uses an end-to-end network, so the accuracy rate under NM # condition is lower than that in the [45].

*2) OULP:* Similar to the settings in [46] and [47], we use CV01 as the train set and CV02 as the test set. The experimental results are shown in Tab. III. The accuracy rates in all comparison cases are shown.

Our method achieves the highest accuracy rate in all gallery views. But when the probe view is 75°, our method is not the best. When the accuracy rate is close to 100%, certain images of very poor quality in the dataset make the randomness of the test results higher and become the decisive factor affecting the accuracy. The method in [7] uses Gait Energy Image (GEI) as the input of the network, which avoids the influence of interfering frames compared with the direct input of original frames in our method, and thus obtains a higher accuracy rate.

TABLE III
AVERAGED RANK-1 ACCURACIES ON **OULP** WITH CROSS CONDITIONS

| Probe view | Method | Gallery view | | | | |
|---|---|---|---|---|---|---|
| | | 55° | 65° | 75° | 85° | mean |
| 55° | Wu *et al.*[10] | 98.8 | 98.3 | 96.0 | 80.5 | 91.6 |
| | He *et al.*[48] | 98.8 | 99.4 | 96.1 | 77.9 | 91.6 |
| | Zhang *et al.*[49] | 99.3 | 99.6 | 99.6 | 99.0 | 99.4 |
| | Ben *et al.*[7] | - | **100** | 99.6 | 98.5 | 99.4 |
| | Ben *et al.*[8] | - | 99.8 | 98.4 | 96.7 | 98.3 |
| | Ours | **99.7** | 99.9 | **100** | **99.4** | **99.8** |
| 65° | Wu *et al.*[10] | 96.3 | 96.3 | 97.3 | 83.3 | 92.3 |
| | He *et al.*[48] | 97.7 | - | 98.5 | 84.4 | 93.5 |
| | Zhang *et al.*[49] | 99.3 | 99.3 | 99.8 | 99.7 | 99.6 |
| | Ben *et al.*[7] | 99.9 | - | **100** | **99.9** | **99.9** |
| | Ben *et al.*[8] | **100** | - | **100** | 99.4 | 99.8 |
| | Ours | 99.7 | **100** | **100** | 99.8 | **99.9** |
| 75° | Wu *et al.*[10] | 94.2 | 97.8 | 98.9 | 85.1 | 92.4 |
| | He *et al.*[48] | 94.8 | 98.9 | - | 86.4 | 93.4 |
| | Zhang *et al.*[49] | 98.8 | 99.4 | 99.6 | 99.6 | 99.3 |
| | Ben *et al.*[7] | 99.6 | **100** | - | **100** | **99.9** |
| | Ben *et al.*[8] | 98.8 | **100** | - | **100** | 99.6 |
| | Ours | **99.7** | 99.8 | **100** | 99.6 | 99.8 |
| 85° | Wu *et al.*[10] | 90.0 | 96.0 | 98.4 | 98.9 | 94.8 |
| | He *et al.*[48] | 86.9 | 97.4 | 99.5 | - | 94.6 |
| | Zhang *et al.*[49] | 98.0 | 99.0 | 98.5 | 99.7 | 98.8 |
| | Ben *et al.*[7] | 98.4 | 99.6 | **100** | - | 99.3 |
| | Ben *et al.*[8] | 96.4 | 99.4 | **100** | - | 98.6 |
| | Ours | **99.8** | **100** | **100** | **99.9** | **99.9** |
| | Ours | **99.7** | **99.9** | **100** | **99.8** | |

## C. Ablation Study

First, we conduct three ablation experiments on the CASIA-B dataset to verify the effectiveness of TPFE and AFRE, and the effectiveness of TPFE and AFPE working together, respectively. Second, we resize the Windows in TPFE and AFRE. Finally, we look forward to the experiments that need to be done in the future.

*1) Effectiveness of PFRE:* Tab. IV shows the impact of each component on RPNet. It is clear that the experiments using

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

TABLE IV

ABLATION STUDY, GROUP A. RESULTS OF MODULES' EFFECT. IT CAN BE SEEN THAT EACH MODULE CAN WORK INDEPENDENTLY OR WITH OTHER MODULES TO OBTAIN A LARGER PERFORMANCE IMPROVEMENT

| Group A | TPFE | AFRE | NM | BG | CL |
|---|---|---|---|---|---|
| a | | | 91.90 | 87.13 | 62.17 |
| b | ✓ | | 94.41 | 89.50 | 73.28 |
| c | | ✓ | 96.32 | 91.01 | 78.63 |
| d | ✓ | ✓ | 96.55 | 92.82 | 80.26 |

TABLE V

ABLATION STUDY, GROUP B. THE RESULT WHEN DIFFERENT WINDOW SIZES ARE SET FOR AFRE. THE WINDOW SIZE OF TPFE IS FIXED DURING THE EXPERIMENT

| Group B | the size of window | NM | BG | CL |
|---|---|---|---|---|
| a | 1 | 96.01 | 91.80 | 77.77 |
| b | 1, 3 | 96.55 | 92.82 | 80.26 |
| c | 1, 5 | 95.98 | 91.22 | 77.49 |
| d | 1, 3, 5 | 96.26 | 91.74 | 78.17 |

TABLE VI

ABLATION STUDY, GROUP C. THE RESULT WHEN DIFFERENT WINDOW SIZES ARE SET FOR TPFE. THE WINDOW SIZE OF AFRE IS FIXED DURING THE EXPERIMENT

| Group C | the size of window | NM | BG | CL |
|---|---|---|---|---|
| a | 1, 4, 16 | 96.36 | 92.27 | 80.48 |
| b | 2, 4, 8 | 96.08 | 91.50 | 78.58 |
| c | 1, 2, 4 | 95.86 | 91.44 | 78.21 |
| d | 4, 8, 16 | 96.67 | 91.79 | 78.13 |
| e | 1, 2, 4, 8, 16 | 96.55 | 92.82 | 80.26 |


Fig. 7. The result when the number of input frames is different. Dotted lines show the trend of the data. More input frames can improve the accuracy under occlusion.

whole PFRE gain better performance than experiment A-a, A-b, and A-c. When we discard both TPFE and AFPE, we get the worst results. These results prove the effectiveness of each part of our method. Each part can bring great improvements to the model.

Seeing Tab. IV longitudinally, we can find that paying attention to the relationship between various parts can significantly improve the recognition accuracy. Among them, the increase in accuracy in the BG # and CL # conditions are higher than the increase in accuracy in the NM # condition. And under NM # condition, the accuracy rate can be close to the highest level with only one sub-module. This indicates that it is more important to focus on the 'parts' in the case of occlusion than it is in the normal case.

*2) The Effect of Different Size Windows in TPFE and AFRE:* In order to get a more intuitive comparison effect, we fix the size of one window when changing the size of another window. The window of TPFE is fixed as [1, 2, 4, 8, 16], and the window of AFRE is fixed as [1, 3]. As shown in Tab. V and Tab. VI, TPFE and AFRE get the best results when the window size is set to [1, 2, 4, 8, 16] and [1, 3], respectively.

Comparing C-c with C-d, we find that a large window is more beneficial to the recognition under normal conditions. Comparing C-a and C-e with C-c and C-d, we find that multi-scale windows are more beneficial to the recognition under strong occlusion.

### D. Quantitative Experiment

*1) The Effect of the Number of Frames Input:* Gait is a biological feature with periodicity. Under normal circumstances, people think that the gait movement in a cycle contains most of gait's characteristics. When the occlusion in the field of view is larger, however, more cycles may bring more comprehensive feature information. Therefore, we conduct adjustment experiments on the number of input frames for training the network, and the experimental results are shown in Fig.7. The experimental results in Fig.7 verify our conjecture: with the increase in the number of input frames during training, the accuracy of the CL state during testing has been improved. However, the accuracy rate in the case of NM has been reduced. We predict that, because the network has a strong learning ability, more inputs will amplify the noise and the network will eventually learn some invalid features.

*2) Generalization Ability:* In order to further illustrate the generalization ability of our model when dealing with large

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

AVERAGED RANK-1 ACCURACIES ON **OUMVLP** WITH CROSS CONDITIONS. ONLY 1500 SUBJECTS WERE USED FOR TRAINING. "TEST SIZE" INDICATES THE NUMBER OF SUBJECTS USED FOR TESTING. "ALL" MEANS TO USE ALL TEST SUBJECTS. "ALL*" MEANS TO USE THE FOUR VIEWS (30°, 45°, 210°, 225°) IN ALL TEST SUBJECTS

| Test Size | 0° | 15° | 30° | 45° | 60° | 75° | 90° | 180° | 195° | 210° | 225° | 240° | 255° | 270° | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 500 | 86.90 | 91.72 | 92.83 | 93.07 | 91.47 | 91.55 | 91.31 | 87.63 | 90.03 | 92.02 | 92.35 | 91.94 | 91.60 | 90.66 | 91.08 |
| 1000 | 84.99 | 90.19 | 92.09 | 92.28 | 90.91 | 91.28 | 90.62 | 86.39 | 89.20 | 91.33 | 91.51 | 90.55 | 90.28 | 89.35 | 90.07 |
| 1500 | 82.59 | 89.39 | 91.75 | 91.98 | 90.52 | 90.47 | 89.69 | 84.84 | 88.40 | 91.02 | 91.36 | 89.84 | 90.19 | 88.75 | 89.34 |
| All | 73.51 | 84.37 | 89.60 | 89.82 | 86.28 | 87.38 | 86.02 | 76.33 | 83.23 | 88.60 | 88.91 | 85.73 | 86.43 | 84.39 | 85.04 |
| All* | - | - | 95.20 | 95.66 | - | - | - | - | - | 94.15 | 94.68 | - | - | - | 94.92 |
| GaitSet[11] | 79.5 | 87.9 | 89.9 | 90.2 | 88.1 | 88.7 | 87.8 | 81.7 | 86.7 | 89.0 | 89.3 | 87.2 | 87.8 | 86.2 | 87.1 |
| GaitPart[14] | 82.6 | 88.9 | 90.8 | 91.0 | 89.7 | 89.9 | 89.5 | 85.2 | 88.1 | 90.0 | 90.1 | 89.0 | 89.1 | 88.2 | 88.7 |
| MvGGAN[45] | 52.6 | 62.8 | 63.9 | 57.5 | 55.4 | 61.3 | 61.9 | 54.8 | 58.8 | 59.3 | 58.5 | 56.6 | 57.5 | 56.8 | 58.4 |

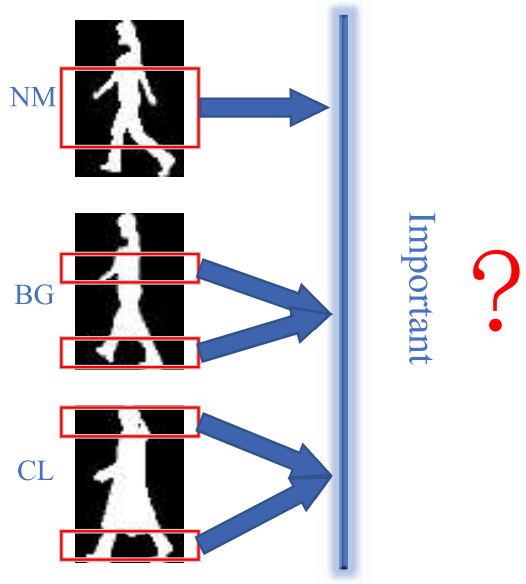



Fig. 8. The part inside the red box indicates the aspect that may play a decisive role in the corresponding situation, and the part outside the red box indicates the aspect that may not be so important.

amounts of data, we conduct relevant experiments on the OUMVLP dataset. Thirty percent of the training data is used to train the model and tests are carried out on test data of different sizes. The experimental results are shown in Tab. VII. "Test Size" indicates the number of subjects used in the model test.

By comparing the data of one to four groups, we can find that our model can maintain high accuracy even when the amount of test data varies greatly. By observing the fifth group of data, we can find that: when the views of 30°, 45°, 225° and 240° are retained, even if all the test subjects (5154) are used, our model can still reach the accuracy rate of 94.9%, which shows the strong generalization ability of our model and the feasibility of gait recognition in practical application.

### *E. Future Outlook*

Gait recognition has undergone many years of research, but its application has not been promoted. The use of deep learning for research improves the recognition accuracy but also creates unexplainable problems, which limits the practical application of gait recognition.

It is pointed out in [49] that different parts of the human body contribute differently to recognition in different situations. As shown in Fig. 8, in case NM, both hands and legs can be used as the basis for identification. In case CL, only the head and feet may play a role in recognition, and other parts will only cause interference. However, there is no clear explanation about which area can play a role or what the proportions of each part should be.

Next, we can do some controlled variable experiments and reach a convincing conclusion through a large amount of experimental data. An interpretable and clear conclusion will promote the development of gait recognition and its application.

## V. CONCLUSION

In this paper, we proposed a method to discover the relationships between the various parts of the gait diagram and build **RPNet** based on it. RPNet consists of two parts: the Convolutional Neural Network (**CNN**) and the Part Feature Relationship Extractor (**PFRE**). PFRE not only extracts features of different scales, but also discover the relationships between local features. Experiments on the **CASIA-B**, **OULP** and **OUMVLP** datasets show that RPNet has excellent performance regardless of occlusion and multiple volumes. Finally, we introduced a qualitative experiment that we performed, discussing the impact of the number of input frames on different walking conditions. The feature aggregation methods in RPNet are all direct splicing. In the future, we will look for better feature aggregation methods. In addition, we will thoroughly study the influence of various parts of the human body on gait recognition in different situations and try to increase the interpretability of the network.

## REFERENCES

[1] L. Wang, T. Tan, H. Ning, and W. Hu, "Silhouette analysis-based gait recognition for human identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 12, pp. 1505–1518, Dec. 2003.

[2] H. M. Thang, V. Q. Viet, N. Dinh Thuc, and D. Choi, "Gait identification using accelerometer on mobile phone," in *Proc. Int. Conf. Control, Autom. Inf. Sci. (ICCAIS)*, Nov. 2012, pp. 344–348.

[3] P. K. Larsen, E. B. Simonsen, and N. Lynnerup, "Gait analysis in forensic medicine," *J. Forensic Sci.*, vol. 53, no. 5, pp. 1149–1153, 2008.

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.

[4] I. Bouchrika, M. Goffredo, J. Carter, and M. Nixon, "On using gait in forensic biometrics," *J. Forensic Sci.*, vol. 56, no. 4, pp. 882–889, Jul. 2011.

[5] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, Dec. 2009, pp. 1–6.

[6] X. Ben, C. Gong, P. Zhang, R. Yan, Q. Wu, and W. Meng, "Coupled bilinear discriminant projection for cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 3, pp. 734–747, Mar. 2020.

[7] X. Ben, C. Gong, P. Zhang, X. Jia, Q. Wu, and W. Meng, "Coupled patch alignment for matching cross-view gaits," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3142–3157, Jun. 2019.

[8] X. Ben, P. Zhang, Z. Lai, R. Yan, X. Zhai, and W. Meng, "A general tensor representation framework for cross-view gait recognition," *Pattern Recognit.*, vol. 90, pp. 87–98, Jun. 2019.

[9] H. Aggarwal and D. K. Vishwakarma, "Covariate conscious approach for gait recognition based upon zernike moment invariants," *IEEE Trans. Cognit. Develop. Syst.*, vol. 10, no. 2, pp. 397–407, Jun. 2018.

[10] 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.

[11] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 8126–8133.

[12] Z. Zhang *et al.*, "Gait recognition via disentangled representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 4710–4719.

[13] T. Wolf, M. Babaee, and G. Rigoll, "Multi-view gait recognition using 3D convolutional neural networks," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 4165–4169.

[14] C. Fan *et al.*, "GaitPart: Temporal part-based model for gait recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14225–14233.

[15] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, Jan. 2021.

[16] D. K. Vishwakarma, R. Kapoor, R. Maheshwari, V. Kapoor, and S. Raman, "Recognition of abnormal human activity using the changes in orientation of silhouette in key frames," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2015, pp. 336–341.

[17] C. Dhiman and D. K. Vishwakarma, "View-invariant deep architecture for human action recognition using two-stream motion and shape temporal dynamics," *IEEE Trans. Image Process.*, vol. 29, pp. 3835–3844, 2020.

[18] D. K. Vishwakarma, "A two-fold transformation model for human action recognition using decisive pose," *Cognit. Syst. Res.*, vol. 61, pp. 1–13, Jun. 2020.

[19] D. K. Vishwakarma and C. Dhiman, "A unified model for human activity recognition using spatial distribution of gradients and difference of Gaussian kernel," *Vis. Comput.*, vol. 35, no. 11, pp. 1595–1613, Nov. 2019.

[20] S. Li, S. Bak, P. Carr, and X. Wang, "Diversity regularized spatiotemporal attention for video-based person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 369–378.

[21] W. Li, X. Zhu, and S. Gong, "Harmonious attention network for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2285–2294.

[22] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, and Q. Tian, "Pose-driven deep convolutional model for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3980–3989.

[23] H. Yao, S. Zhang, R. Hong, Y. Zhang, C. Xu, and Q. Tian, "Deep representation learning with part loss for person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2860–2871, Jun. 2019.

[24] P. Zhang, J. Xu, Q. Wu, Y. Huang, and X. Ben, "Learning spatial-temporal representations over walking tracklet for long-term person re-identification in the wild," *IEEE Trans. Multimedia*, early access, Oct. 6, 2020, doi: 10.1109/TMM.2020.3028461.

[25] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 441–444.

[26] H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, "The OU-ISIR gait database comprising the large population dataset and performance evaluation of gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

[27] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–14, Dec. 2018.

[28] F. Jean, A. B. Albu, and R. Bergevin, "Towards view-invariant gait modeling: Computing view-normalized body part trajectories," *Pattern Recognit.*, vol. 42, no. 11, pp. 2936–2949, Nov. 2009.

[29] W. Lu, W. Zong, W. Xing, and E. Bao, "Gait recognition based on joint distribution of motion angles," *J. Vis. Lang. Comput.*, vol. 25, no. 6, pp. 754–763, Dec. 2014.

[30] M. Deng and C. Wang, "Human gait recognition based on deterministic learning and data stream of microsoft kinect," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3636–3645, Dec. 2019.

[31] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[32] R. A. Guler, N. Neverova, and I. Kokkinos, "DensePose: Dense human pose estimation in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[33] 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.

[34] E. Zhang, Y. Zhao, and W. Xiong, "Active energy image plus 2DLPP for gait recognition," *Signal Process.*, vol. 90, no. 7, pp. 2295–2302, Jul. 2010.

[35] K. Bashir, T. Xiang, and S. Gong, "Gait recognition without subject cooperation," *Pattern Recognit. Lett.*, vol. 31, no. 13, pp. 2052–2060, Oct. 2010.

[36] C. Chen, J. Liang, and X. Zhu, "Gait recognition based on improved dynamic Bayesian networks," *Pattern Recognit.*, vol. 44, no. 4, pp. 988–995, Apr. 2011.

[37] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On Input/Output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019.

[38] Y. Zhou, Y. Huang, Q. Hu, and L. Wang, "Kernel-based semantic hashing for gait retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 2742–2752, Oct. 2018.

[39] R. Liao, C. Cao, E. B. Garcia, S. Yu, and Y. Huang, "Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations," in *Proc. Chin. Conf. Biometric Recognit.*, 2017, pp. 474–483.

[40] Y. Fu *et al.*, "Horizontal pyramid matching for person re-identification," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, pp. 8295–8302, Jul. 2019.

[41] Y. Sun, L. Zheng, Y. Yang, and S. Wang, "Beyond part models: Person retrieval with refined part pooling," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 480–496.

[42] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 4321–4329.

[43] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, *arXiv:1703.07737*. [Online]. Available: http://arxiv.org/abs/1703.07737

[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[45] X. Chen, X. Luo, J. Weng, W. Luo, H. Li, and Q. Tian, "Multi-view gait image generation for cross-view gait recognition," *IEEE Trans. Image Process.*, vol. 30, pp. 3041–3055, 2021.

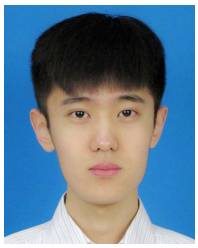
[46] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1602–1615, Jul. 2016.

[47] D. Muramatsu, Y. Makihara, and Y. Yagi, "Cross-view gait recognition by fusion of multiple transformation consistency measures," *IET Biometrics*, vol. 4, no. 2, pp. 62–73, Jun. 2015.

[48] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 102–113, Jan. 2019.

[49] 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.


Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.



**Hao Qin** was born in Shandong, China, in 1998. He received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2020, where he is currently pursuing the M.S. degree in control science and engineering. His research interests include machine learning, deep learning, and gait recognition.

**Zhenxue Chen** was born in Shandong, China, in 1977. He received the B.S. degree in automatic from the School of Electrical Engineering and Automation, Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from the School of Information Science and Engineering, Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, in 2007. From 2012 to 2013, he was a Visiting Scholar with Michigan State University, East Lansing, USA. He is currently a Professor with the School of Control Science and Engineering, Shandong University. He has published over 100 papers in refereed international leading journals/conferences, such as IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS (TII), IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY (TCSVT), IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (TIFS), IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY (TVT), IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS (TITS), *Information Sciences*, *Neurocomputing*, *Neural Computing and Applications*, and *SP-IC*. His main research interests include image processing, pattern recognition, and computer vision, with applications to face recognition.

**Qingqiang Guo** was born in Shandong, China, in 1971. He received the Ph.D. degree in control science and engineering from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2010. He is currently an Associate Professor with the School of Control Science and Engineering, Shandong University. His research interests include production scheduling, process control and optimization, system analyze and identification, and computer vision.

**Q. M. Jonathan Wu** (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, U.K., in 1990. In 1995, he was with the National Research Council of Canada for ten years, where he became a Senior Research Officer and the Group Leader. He is currently a Professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 300 peer-reviewed articles in computer vision, image processing, intelligent systems, robotics, and integrated microsystems. His current research interests include machine learning, 3D computer vision, video content analysis, interactive multimedia, sensor analysis and fusion, and visual sensor networks. He holds the Tier 1 Canada Research Chair in automotive sensors and information systems. He was an Associate Editor of IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART A: SYSTEMS AND HUMANS, and the *International Journal of Robotics and Automation*. He is an Associate Editor of the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS and the journal of *Cognitive Computation*. He has served on technical program committees and international advisory committees for many prestigious conferences.

**Mengxu Lu** was born in Jiangsu, China, in 1997. She received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2019, where she is currently pursuing the M.S. degree in control science and engineering. Her research interests include machine learning, deep learning, and semantic segmentation.

Authorized licensed use limited to: Zhejiang University. Downloaded on June 22,2026 at 12:06:10 UTC from IEEE Xplore. Restrictions apply.