Received: 6 February 2021 | Revised: 15 March 2021 | Accepted: 14 April 2021

CAAI Transactions on Intelligence Technology

DOI: 10.1049/cit2.12051


ORIGINAL RESEARCH PAPER

# Learning discriminative representation with global and fine-grained features for cross-view gait recognition

Jing Xiao[1] | Huan Yang[1] | Kun Xie[1] | Jia Zhu[2] | Ji Zhang[3]

[1]School of Computer Science, South China Normal University, Guangzhou, Guangdong, China

[2]Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, Zhejiang, China

[3]School of Sciences, University of Southern Queensland, Toowoomba Qld, Australia

**Correspondence**

Jing Xiao, School of Computer Science, South China Normal University, Guangzhou, Guangdong, 510631, China.
Email: xiaojing@scnu.edu.cn
Jia Zhu, Key Laboratory of Intelligent Education Technology and Application of Zhejiang Province, Zhejiang Normal University, Jinhua, Zhejiang, 321004, China.
Email: jzhu@m.scnu.edu.cn

**Funding information**

Key Area Research and Development Program of Guangdong Province, Grant/Award Number: 2019B111101001; Natural Science Foundation of Guangdong Province, Grant/Award Number: 2018A030313318

## Abstract

In this study, we examine the cross-view gait recognition problem. Many existing methods establish global feature representation based on the whole human body shape. However, they ignore some important details of different parts of the human body. In the latest literature, positioning partial regions to learn fine-grained features has been verified to be effective in human identification. But they only consider coarse fine-grained features and ignore the relationship between neighboring regions. Taken the above insights together, we propose a novel model called GaitGP, which learns both important details through fine-grained features and the relationship between neighboring regions through global features. Our GaitGP model mainly consists of the following two aspects. First, we propose a Channel-Attention Feature Extractor (CAFE) to extract the global features, which aggregates the channel-level attention to enhance the spatial information in a novel convolutional component. Second, we present the Global and Partial Feature Combiner (GPFC) to learn different fine-grained features, and combine them with the global features extracted by the CAFE to obtain the relevant information between neighboring regions. Experimental results on the CASIA gait recognition dataset B (CASIA-B), The OU-ISIR gait database, multi-view large population dataset, and The OU-ISIR gait database gait datasets show that our method is superior to the state-of-the-art cross-view gait recognition methods.

## 1 | INTRODUCTION

Gait recognition is a promising video-based biometric identification technology applied to identify individuals by their walking patterns. Compared to other biometric technologies, such as the face, fingerprint and iris recognition, gait recognition has the advantages of non-contact, long-distance and no explicit cooperative interest-subjects. Therefore, gait recognition has a potentially wide range of applications in video surveillance. As the accuracy increases, gait recognition technology will definitely become another effective tool for crime prevention, forensic identification and social security. In order to improve the accuracy of recognition, we need to overcome various external factors, including walking speed, bag-carrying, coat-wearing and camera viewpoint, that cause dramatic changes in gait appearance. As shown in Figure 1, the appearance of gait walking changes observably in different directions, which may result in the similarity of inter-class common attributes greater than that of intra-class common attributes, and brings challenges to gait recognition.

There are several attempts in the literature to solve the cross-view gait recognition problem. A common strategy is to extract global features by treating the whole human image as a unit. It is worth mentioning that many methods [1–4] use attention mechanisms to improve the performance of the model, and our method is no exception. However, due to the diversity of gait walking conditions in the cross-view situation, some important details are often ignored in the global features.

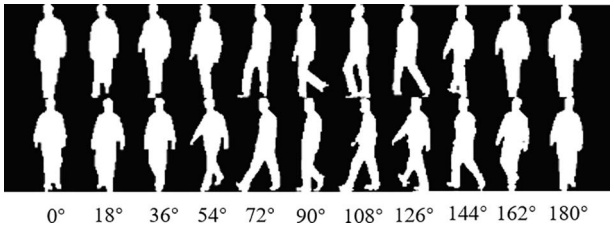


**FIGURE 1** From left to right are silhouettes of all views in The CASIA gait recognition dataset B (CASIA-B), gait dataset, which possess evidently different shapes and moving patterns during walking

Another learning strategy considers that different parts of the human body poss evidently various shapes and moving patterns during walking [5–10]. They aim to learn fine-grained features through specific regions. Unfortunately, they only consider coarse fine-grained features and ignore the relationship between neighboring regions. To solve the problems in the above two strategies, we propose a novel model called GaitGP, which learns both important details through fine-grained features and the relationship between neighboring regions through global feature representation.

Our novel model GaitGP consists of the following two components. The first component is a Channel-Attention Feature Extractor (CAFE), which is a novel application of convolution and can extract global features with channel attention mechanism. The other one is the Global and Partial Feature Combiner (GPFC), which learns fine-grained features in specific regions of images. Moreover, the GPFC combines the global features extracted by the CAFE with fine-grained features to obtain relevant information between neighboring regions.

In the CAFE, we jointly learn attention selection and feature representation to extract global features with a channel attention mechanism. From the experimental data, we find that channel attention does enhance the performance of the model compared to the original global features extracted from the image. Therefore, an effective channel attention mechanism method, called the Channel-level Spatial Pooling (CSP) is introduced to select the channel attention information and optimize the global features. Additionally, in order to improve the compatibility between channel attention selection and global features, our novel convolution layer adopts the partitionable stacking design, which will be discussed specifically in Section 3.2.

In the GPFC, we divide the global feature map extracted from the CAFE into several sub-branches. To combine the global feature and the fine-grained features, the first sub-branch contains only one whole partition to preserve the global information. In the remaining sub-branches, we divide the global feature maps into different numbers of stripes as part regions to learn local feature representations independently [5]. More details will be discussed specifically in Section 3.3.

More simply, we summarise our contributions as follows:

- We propose a novel model called the GaitGP, which learns both important details through fine-grained features and the relationship between neighboring regions through global feature representation.
- We propose a CAFE for the optimization of global feature representation.
- We propose a GPFC for combining the global and fine-grained features.
- For gait recognition accuracy, we combine the above aspects to conduct a large number of experimental ablation experiments on the widely used gait datasets the CASIA gait recognition dataset B (CASIA-B) [11], The OU-ISIR gait database, multi-view large population dataset (OU-MVLP) [12] and The OU-ISIR gait database (OULP) [13]. Compared to several state-of-the-art methods, GaitGP shows superiority.

## 2 | RELATED WORK

### 2.1 | Cross-view gait recognition

To adapt to the situation of cross-view for gait recognition, one of the most typical gait recognition methods is treating the whole human body shape as a unit to extract features and can be divided into two categories: model-based [14–17] and appearance-based [18–25]. The model-based method tries to reconstruct the human 3D-body and motion models to identify individuals. Wolf et al. [14] used to model the dynamic characteristics of the gait sequence to express the overall understanding of the gait sequence. The gait silhouettes under different views are mapped on a common template by the 3D-model, but it is difficult to train because of the complexity of network architecture.

Many appearance-based methods in this fashion perform gait recognition in a more lightweight (easily to train) network architecture. Inspired by the great achievements in face recognition and action recognition, some researchers leverage generative methods to reconstruct the gait template in all views. The generative adversarial network (GAN) [26] is used to generate invariant side-view gait images to adapt to the situation of appearance changes caused by different clothing. Yu et al. [22] proposed a unified cross-view gait recognition model based on a generative framework to learn view-invariant features. A multi-loss strategy is used in GaitGAN [27] to optimize the network to increase the inter-class distance and reduce the intra-class distance. All these methods compress the gait silhouettes from different views into a uniform template for gait recognition. However, it is believed that these methods retain unnecessary sequential constraints for periodic gait [3] and ignore some important details of different parts of the human body.

For learning more detailed information to enhance feature representation, many advanced methods in Re-Identification (Re-ID) task [5–9, 28–30] have proved that locating important body parts from images to represent local identity information is an effective method to improve the accuracy and robustness of recognition. One of the most commonly used strategies is to split the feature map into strips and merge them into column vectors. Wang et al. [31] designed a Multiple Granularity Network with multiple branches, which uniformly partitions the images into several stripes, and varies the number of parts in different local branches to obtain local feature representation with multiple

granularities. Fu et al. [8] proposed a simple and effective horizontal pyramid matching method to fully exploit various partial information of a given person. In the task of gait recognition, many of the latest articles have applied the strategies of fine-grained features. Chao et al. [3] used Horizontal Pyramid Mapping to map the set-level feature into a more discriminative space for robust feature representation. Zhang et al. [32] employed the idea of part-based unified segmentation to extract local features of gait. However, these methods only consider coarse fine-grained attentional features and ignore the relationship between neighboring regions.

To learn both the complement of important details through fine-grained features and the relationship between neighboring regions, in this work, the proposed model GaitGP combines the attention information to learn global feature representation and aggregate it with the fine-grained features to make feature representation more robust.

## 2.2 | Deep learning on attention

One common learning strategy is long short term memory (LSTM)- based [32, 33]. They employ LSTM to temporal attention scores to pay more attention on those discriminative frames, and thus, improving the overall performance. However, these methods are considered to retain the unnecessary sequence constraints on the periodic gait. Additionally, some new approaches in the Re-ID task combine the local attention-based representation of the image to improve performance [1, 30, 34–38]. Li et al. [34] proposed a Spatial Transformer Network (STN) with spatial constraints [36] to locate deformable pedestrian features. Zhao et al. [39] built a hard attention model by the STN to search for components, given the pre-defined spatial constraints. Li et al. [1] presented Harmonious Attention convolutional neural network (CNN) for joint learning of different levels of visual attention subject along with simultaneous optimization of feature representation.

Inspired by the successful application of visual attention, some methods directly perform on random sequences to get attention information, thus, avoiding unnecessary sequence constraints on the periodic gait. GaitSet [3] presented the Set Pooling applying attention mechanism [1, 40, 41] to improve its performance. GaitPart [4] applied the channel-wise attention mechanism [6, 42, 43] to the re-weighted micro-motion feature, which aims to overcome the limitation of the global feature. The above methods show that the attention information is beneficial to improve the performance of gait recognition. Therefore, in our method, we propose the CSP to learn channel-wise attention to enhance the global feature.

# 3 | PROPOSED METHOD

In this section, we first summarize the overall network architecture of the GaitGP model. This is followed by a detailed description of the two components of the model, that is, the CAFE and GPFC.

## 3.1 | Overall framework

The overall of the proposed method is shown in Figure 2. Given a dataset of $n$ people with identity $y_i$, $i \in 1, 2, \ldots, n$, we assume that the sequence of each identity is $X_i$. $s$ silhouettes given from each $X_i$ are expressed as $X_i = \{x_i^j, j = 1, 2, \ldots, s\}$. We first use the CAFE to jointly perform attention selection and feature representation. Then, the global features are extracted through a channel attention mechanism, which is formulated as follows:

$$\chi_\tau = CAFE(\varphi(X_i)) \tag{1}$$

where $\chi_\tau$ denotes the output feature map of the CAFE and $\varphi$ denotes the function of attention selection, which is implemented by the CSP in the CAFE. The details of the CSP will be introduced in Section 3.2.

Then, the GPFC divides $\chi_\tau$ into $t$ sub-branches. Each sub-branch is horizontally split into $p = 2^\gamma$, $\gamma = 1, 2, \ldots$, denoted as $\chi_\tau^p$ partitions, Finally, the GPFC combines all the fine-grained features and the global feature $\chi_\tau$ to learn the relationship between the neighbor regions. The GPFC is formulated as follows:

$$\nu_\delta = GPFC\left(\sum_{i=1}^{t} \delta\left(\chi_\tau^p\right)\right) \tag{2}$$

where $\nu_\delta$ denotes the column vector down-sampled by the $\delta$; $\delta$ denotes a Multi-Granularity Mapping (MGM) module. More details will be introduced in Section 3.3.

Finally, we choose the separate triplet loss [3, 4] to train the proposed model.

## 3.2 | Channel-Attention Fusion Extractor

The CAFE learns global features with channel-level attention to enhance representation. There are two components in the CAFE: The CSP which aims to learn the attention information and the Partitionable Convolution layer (PConv) used to extract the global feature for integrating the attention information freom the CSP. Next, the CSP is described in detail first, followed by the exact structure of the PConv.

### 3.2.1 | Channel-level Spatial Pooling

To enhance the expressiveness of global features, the CSP learns a channel-wise attention map to refine it. As shown in Figure 2, each block of the Partial Branch contains a CSP, assuming that $f_b \in \mathbb{R}^{c \times s \times h \times w}$ is the input feature map of the CSP; $b$ represents the block in the CAFE; $c$ is the number of channels; $s$ is the length of the gait sequence and $(h, w)$ is the size of each feature map.

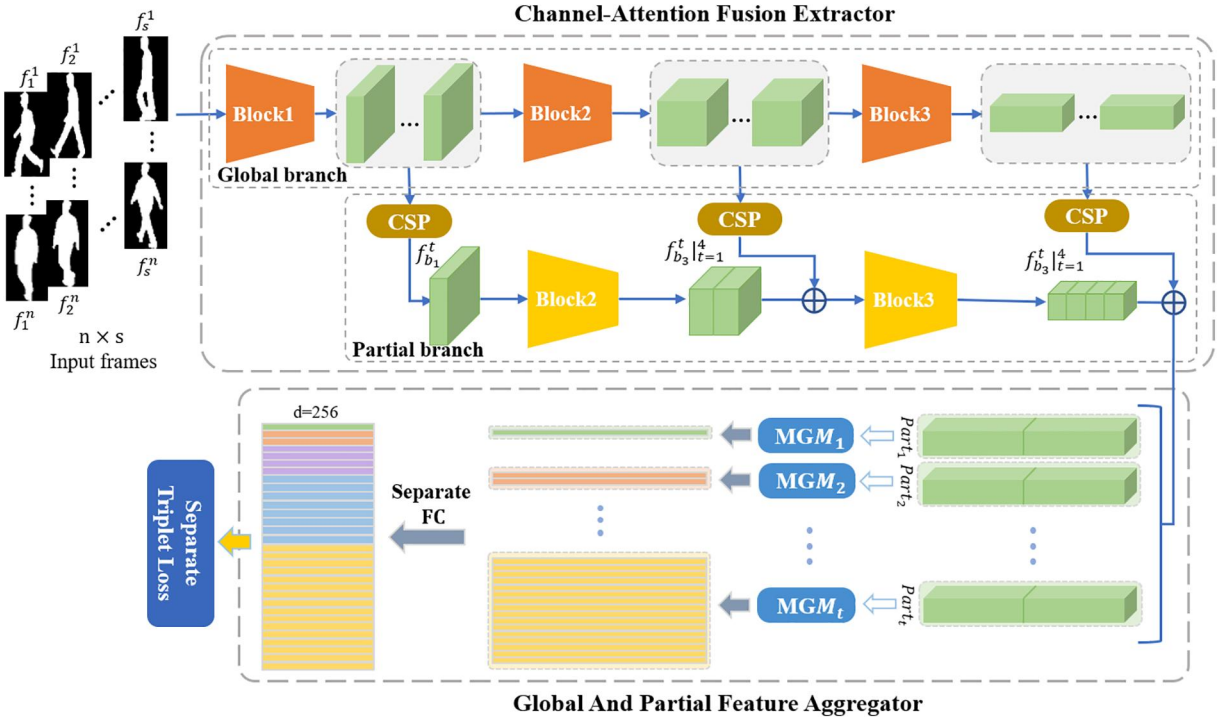


**FIGURE 2** The framework of GaitGP. Channel-Attention Fusion Extractor is consistsof CSP and Blocks. CSP represents the Channel-level Spatial Pooling and the Blocks are composed of two convolutional units (PConvs). In the Global Branch, PConv is mainly utilized to extract global features; while in the Partial Branch, PConv is used to collect channel-level attention information. Global And Partial Feature Aggregator is used to gather the global and fine-grained features. MGM represents the Multi-Granularity Mapping. Note that the MGMs are independent, each of which has a different scale. The dimension of the final feature is 256. FC, Fully Convolution

Since the length of the input gait may be different, many previous works [3, 4, 10] successfully utilize pooling to aggregate the gait information of elements in a sequence. Therefore, as shown in Figure 3, we first use Spatial Pooling to aggregate the information of gait elements to represent the gait motion pattern. A natural choice of the Spatial Pooling is to apply the statistical max function [3] on channel dimension. We pre-divide $f_b$ into $\tau$, $\tau \in [1, 2, 4, \ldots]$ channel-level partitions to aggregate the information of gait elements which is formulated as follows:

$$f_{sp}^{\tau} = max_s\left(\frac{1}{H \times W}\sum_{h=1}^{H}\sum_{w=1}^{W} f_{b_{\frac{c}{\tau},s,h,w}}\right) \tag{3}$$

where $f_{b_{\frac{c}{\tau},s,h,w}} \in \mathbb{R}^{\frac{c}{\tau}\times s\times h\times w}$ is the sequence-level feature map and $f_{sp}^{\tau} \in \mathbb{R}^{c\times h\times w}$ is the frame-level feature map.

Then, $f_{sp}^{\tau}$ is divided into two streams denoted as $f_{sp^1}^{\tau}$ and $f_{sp^2}^{\tau}$. $f_{sp^1}^{\tau}$ is used to extract spatial attention by the ConvNet. The ConvNet consists of $1 \times 1$ convolutional layer. We assume that each pre-partitioned $\tau$ partition corresponds to a ConvNet module for extracting the local spatial attention. Note that these ConvNet modules are independent. Then, we get a saliency channel-level attention score $f_{score}$, formulated as follows:

$$f_{score} = Concat\sum_{1}^{\tau} ConvNet\left(f_{sp^1}^{\tau}\right) \tag{4}$$

where *Concat* represents the concatenation on the dimension of the channel.

Finally, $f_{score}$ is merged into the $f_{sp}^{2}$ collected by the statistical functions, formulated as follows:

$$f_{weight} = SP\left(f_{sp^2}^{\tau}\right) \oplus f_{score} \tag{5}$$

where $f_{weight} \in \mathbb{R}^{c\times h\times w}$ is the final output of the CSP and $\oplus$ is a channel-wise fusion operation. $f_{weight}$ contains the frame-level Global information.

### 3.2.2 | Partitionable convolution layer

The PConv is a basic unit of the blocks in the CAFE. To improve the compatibility between attention information and the global feature, the PConv is designed to be partitionable. As shown in Figure 2, we design the CAFE as a multi-branched structure. In the Global Branch, the PConv is mainly utilised to extract global features; while in the Partial Branch, the PConv is used to collect channel-level attention information. The global feature extracted in different blocks of

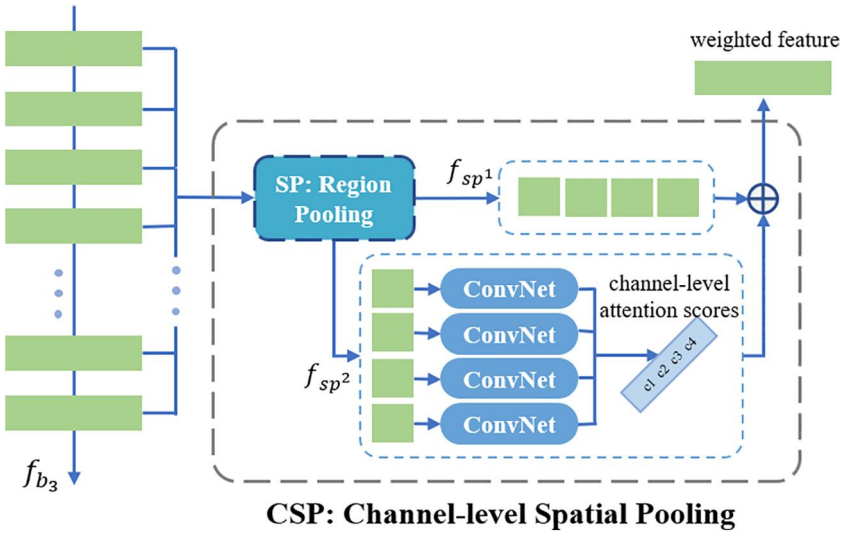


**FIGURE 3** The structure of CSP. Take $\tau = 4$ as an example. The SP module applies an improved statistical max function to gather the most discriminative feature. The ConvNet is a convolutional layer with an activation function rectified linear unit (ReLU), which obtains the channel-level attention scores. CSP, Channel-level Spatial Pooling

the Global Branch are added to the Partial Branch. In order to adapt to the various-level fusion of global features and channel-level attention information, we pre-define that each block has $\tau$ channel-wise regions. In the initialization block (Block1), the PConv ($\tau = 1$) is equivalent to the regular convolutional layer. In the remaining blocks (Block2 and Block3), the PConvs are divided; The input global features are divided into t channel regions for convolution operation, and then vertically spliced together as the final output.

Supposing the output of the Global Branch is $S_{global} \in \mathbb{R}^{c\times h\times w}$ and output of partial branch is $S_{part} \in \mathbb{R}^{c\times h\times w}$, we connect both the two feature maps, represented as follows:

$$S_{part} = CSP(f_i) \oplus PConv_p\left(f_{weight}\right) \tag{6}$$

$$S_{global} = PConv_g(f_i) \tag{7}$$

where $PConv_p$ and $PConv_g$ represent the convolutional layer in the Partial Branch and in the Global Branch, respectively. $\oplus$ denotes the concatenate operation.

$$S_{cafe} = Concat\left(\frac{S_{global}}{S_{part}}\right) \in \mathbb{R}^{2c\times h\times w} \tag{8}$$

where $S_{cafe}$ is the final feature map of the CAFE, *Concat* represents the function Concatenate. Note that $2c$ means that the dimension of the channel becomes twice after the operation *Concat*.

Different layers have different receptive fields and each block contains two PConv layers, as shown in Figure 4(b). The exact structure and parameters of each PConv are shown in Table 1. As shown in Figure 4(a), taking the PConv in Block3 as an example, the input feature map is horizontally divided into $\tau = 4$ partitions, which are operated independently. Then, the obtained channel-level feature vectors are spliced vertically as the final output.

## 3.3 | Global and Partial Feature Aggregator

In literature, splitting the feature map into strips is commonly used in person Re-ID task [5, 8, 31]. Horizontal Pyramid Pooling (HPP) [8] proposes to learn different fine-grained features with four scales, and thus, can help the deep network focus on features with different sizes to gather both partial and global information. We improve the HPP to obtain relevant information between neighboring regions. The most obvious modification is that we divide the subsequent part into five independent sub-branches after the CAFE process. Each sub-branch has similar architecture with different scales.

Specifically, the GPFC has $\rho$ scales. On scale $\rho$, the feature map, $S_{cafe}$, extracted by the CAFE is split into five independent sub-branches, expressed as $\{Part_t\}_{t=1,2,\ldots,5}$. Each sub-branch uses an MGM module with different scales, as shown in Figure 2. The MGM splits each $Part_t$ into $\rho = 2^{m-1}$ on height dimension, that is, $\sum_{m=1}^{M} 2^{M-1}$ strips in total, where $m \in \{1, 2, \ldots, M\}$, M = 1. The upper sub-branch $Part_1$ contains only one whole partition (preserve global feature), which is used to supplement the relevant information between neighboring regions of other sub-branches. For the remaining four sub-branches, the $S_{cafe}$ is split into different $\rho$ scales, that is, horizontally divided into $\rho$ stripes to learn different fine-grained features independently.

Moreover, the structure of the MGM module is shown in Figure 5. On scale $\rho$, the Separate Max Pooling (SMP) is applied to downsample $S_{cafe}$ into 3-D strip features of equal size. Then the Separate Conv1dNet (SC) module is leveraged to reduce the dimension, presented as $v_t$, which consists of a 1-D convolutional layer with a kernel size of $\rho$. The specific parameters of each MGM component are shown in Table 2. The MGM is formulated as follows:

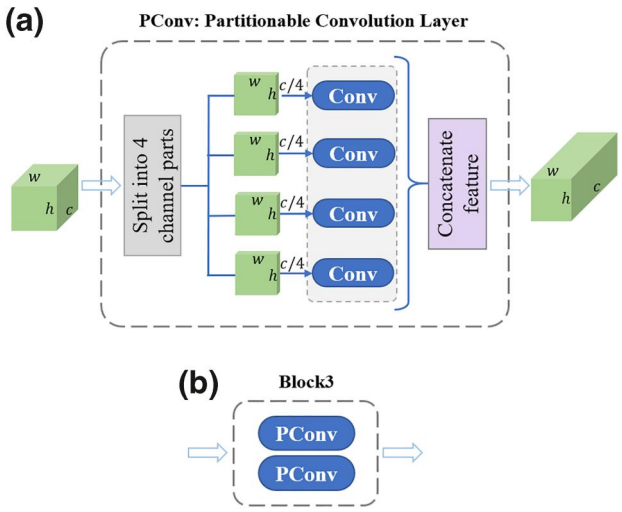


**FIGURE 4** (a) The illustration of the PConv in Block3 and the dimension of the input feature map is expressed as $c \times h \ w$. (b) Block3 is a deep-layer block and consists of two PConvs. PConv, Partitionable Convolution

**TABLE 1** The exact structure of the CAFE and the specific parameters of PConv. In_D, Out_D, Kernel represent the input dimension, output dimension and kernel size of the PConv, respectively. In particular, $\tau$ indicates the pre-defined partition in the PConv. Feature denotes the output feature maps of each block

| Block | Layer | $\tau$ | In_D | Out_D | Kernel | Feature |
|---|---|---|---|---|---|---|
| Block1 | PConv1 | 1 | 1 | 32 | $(5 \times 5, 2)$ | $f_1^{b1}$ |
| | PConv2 | 1 | 32 | 32 | $(3 \times 3, 1)$ | |
| | MaxPool2d, kernel size = 2, stride = 1 | | | | | |
| Block2 | PConv3 | 2 | 32 | 64 | $(3 \times 3, 1)$ | $\{f_t^{b2} \mid_{t=1}^{T}\}$ |
| | PConv4 | 2 | 64 | 64 | $(3 \times 3, 1)$ | |
| | MaxPool2d, kernel size = 2, stride = 1 | | | | | |
| Block3 | PConv5 | 4 | 64 | 128 | $(3 \times 3, 1)$ | $\{f_t^{b3} \mid_{t=1}^{T}\}$ |
| | PConv6 | 4 | 128 | 128 | $(3 \times 3, 1)$ | |
| | MaxPool2d, kernel size = 2, stride = 1 | | | | | |

Abbreviations: CAFE, Channel-Attention Feature Extractor; PConv, Partitionable Convolution.

$$GPFC(v_t) = \sum_{i=1}^{t} MGM_t(S_{cafe}) \tag{9}$$

$$MGM_t(S_{cafe}) = Concat(SC(v^p)) \tag{10}$$

$$v^p = SMP\left(\sum_{p=1}^{5}\sum_{j=1}^{\rho} v_j^p\right) \tag{11}$$

where $MGM_t$ is a $\rho$-granularities extraction feature; $v_j^p$ is horizontally divided into $\rho$ scales; $v_t$ is the aggregated output vector and *Concat* represents the concatenate operation. The SMP is implemented by the 1-D Max Pooling with the kernel size of $\rho$, which is formulated as follows:

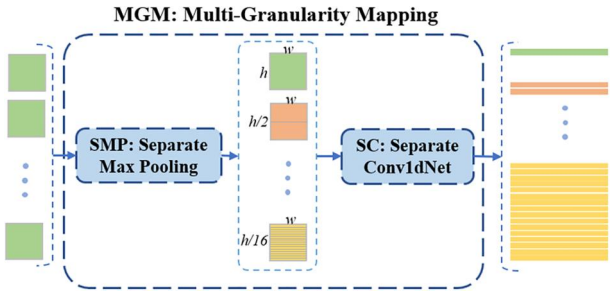


**FIGURE 5** The structure of Multi-Granularity Mapping (MGM). Take scale $\rho = 16$ as an example, the Separate Max Pooling (SMP) is applied to downsample. The Separate Conv1dNet (SC) is leveraged to reduce the dimension

$$SMP(v_j^p) = Maxpool2d(v_j^p) \tag{12}$$

Finally, we perform the Separate Fully Convolution (FC) layer to obtain the final features of the GaitGP, described as *fc*, formulated as follows:

$$v_{fc} = Sepatate\ FC\left(\sum_{j=1}^{\rho} v_t\right) \tag{13}$$

In the testing phase, to obtain the discriminating ability, we splice all the features down to 256 dimensions as the final feature map, combining the global and fine-grained information to improve the comprehensiveness of the learning features.

## 3.4 | Implementation details

### 3.4.1 | Loss function

As shown in Figure 2, we add a Separated Triplet Loss function to supervise learning, which applies the *Separate Batch All (BA+) triplet loss* [44] to train the network and use the corresponding column feature vectors between the different adversarial samples to calculate the loss. The triplet loss is defined as follows:

$$L_{tri} = \{D(N_\gamma, N_p) - D(N_\gamma, N_n) + m\}_+ \tag{14}$$

where $N_\gamma$ is a random sample. $N_p$ is a positive sample with the same identity as the $N_\gamma$. $N_n$ is a negative sample with a different identity from the $N_\gamma$. $m$ is the margin of the triplet loss. The operation$[\vartheta]_+$ is equal to $max(\vartheta, 0)$.

### 3.4.2 | Training

The input of the network is a series of silhouettes. We randomly select samples from the entire gait sequence, which can be regarded as a time data enhancement method. We sample a batch of size $n \times s$ from the training set, where $n$

**TABLE 2** Comparison of the settings for the MGM in five sub-branches. "Sub" refers to the name of sub-branches. "P" refers to the number of partitions on feature maps. "Map Size" refers to the size of the output feature maps from each branch. "Dim" refers to the dimensionality and number of features for the output representations. "Feature" means the symbols for the output feature representation

| Sub | $\rho$ | Map Size | Dim | Feature |
|---|---|---|---|---|
| $Part_1$ | 1 | $(8 \times 1)$ | $256 \times (2 \times 16)$ | $v_{j=1}^{p1}$ |
| $Part_2$ | 2 | $(8 \times 1)$ | $256 \times (2 \times 16)$ | $v_j^{p2}\|_{j=1}^{2}$ |
| $Part_3$ | 4 | $(4 \times 1)$ | $256 \times (4 \times 16)$ | $v_j^{p3}\|_{j=1}^{4}$ |
| $Part_4$ | 8 | $(2 \times 1)$ | $256 \times (8 \times 16)$ | $v_j^{p4}\|_{j=1}^{8}$ |
| $Part_5$ | 16 | $(1 \times 1)$ | $256 \times (16 \times 16)$ | $v_j^{p5}\|_{j=1}^{16}$ |

Abbreviation: MGM, Multi-Granularity Mapping.

represents the number of people with different *id*s, and *s* represents the number of different sequences used by each person with the same *id* in the batch. Sampling strategies in [3, 4] are applied, and the *Separate Batch All (BA+) triplet loss* [44] is used to calculate the loss.

#### 3.4.3 | Testing

The gait sequence is tested using the spatio-temporal features extracted for each gait sequence. The average Euclidean distance between the gallery and the feature column vector of the gallery can be used to match the metric.

## 4 | EXPERIMENTS

In this section, we first describe two databases, CASIA-B and OU-MVLP, to evaluate our model GaitGP, followed by comparing the performance of GaitGP with the state-of-the-art methods and ending with ablation study on CASIA-B to verify the effectiveness of each component in GaitGP.

### 4.1 | Datasets and training details

#### 4.1.1 | CASIA-B

CASIA-B [11] is a widely used gait dataset containing 124 subjects, each of which includes 11 views. Among the views, there are 10 sequences with three gait conditions; one normal condition normal (NM) that includes six sequences. The first 4 sequences NM#01-04 form a gallery, and the remaining two sequences NM#05-06 are used as probes. In addition to the normal condition sequence, there are two sequences; one is wearing a coat cloth (CL)#01-02, the other is carrying a bag (BG)#01-02. The dataset enables researchers to simultaneously study cross-view and cross-wearing issues, in other words, each body contains $11 \times (6 + 2 + 2) = 110$ sequences. There are various experimental schemes [45] based on CASIA-B to verify the feasibility and effectiveness of the proposed method. For fairness, this study strictly follows the popular protocol [6]. Besides, there are three training settings which are configured according to the different training scales in the training stage [3], that is, small-scale training (ST), medium-scale training (MT), and large-scale training (LT). Among them, 124 subjects are divided into two groups; 24, 63, and 74 subjects are put into the training set, and the remaining subjects are reserved for testing. During the test, the first 4 sequence conditions of NM (NM#01-04) are regarded as a gallery and the rest are divided into three subsets of walking conditions based on these six sequences, which are the NM subset of NM#05-06, the other BG subset of BG#01-02, and the last CL subset of CL#01-02.

#### 4.1.2 | OU-MVLP

OU-MVLP [12] is the newly released public gait database with the largest view changes, which consists of 10,307 subjects; each subject containing 14 views $(0, 15, \ldots, 90; 180, 195, \ldots, 270)$. We use the first 5153 for training and the remaining 5154 for testing. There are two sequences in the dataset. In the testing stage, the sequence #01 is classified as the gallery set, and the other sequences #00 are classified as the probe set. According to [12], four typical viewing angles $(0°, 30°, 60°, 90°)$ are evaluated. In addition to doing these four typical views, we conduct experiments with all the views [3, 4, 32, 46]. The data set can provide us with stable comparison results.

#### 4.1.3 | OULP

OULP [13] is a large dataset with only 4 view angles(55°, 65°, 75°, 85°). There are 4,007 subjects (2135 males and 1872 females) with ages ranging from 1 to 94 years and each subject containing two sequences, one in the gallery and the other as a probe sample. Compared with CASIA-B, OULP has smaller view differences and no variants in walking conditions. However, the large number of subjects enables us to compare different gait recognition approaches with statistical significance. Our experimental setting is the same as in [20], since not all samples of each subject are covered from four view angles. A total of 3714 subjects (according to the file of first view angle) are used in the subsequent experiments. We use 1857 subjects as the train set and the rest as the test set. Note that the original silhouettes have already been cropped and aligned. We directly use the given silhouettes to construct the gait templates.

#### 4.1.4 | Training details

During the experiment, the length *s* of the input gait sequences is set to 30, the same as [3, 4]. We use the method mentioned in [12] to crop, align all input sequences, and adjust their size to

**TABLE 3** In three experimental environments with different sample sizes(BT, MT, LT), CASIA-B's average level 1 accuracy under all viewing angles and different conditions (not including the same viewing angle)

| Size | Probe | Gallery NM#1-4 | 0°-180° | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | Mean |
| ST(24) | NM#5-6 | CNN-LB [20] | 54.8 | - | - | 77.8 | - | 64.9 | - | 76.1 | - | - | - | - |
| | | GaitSet [3] | 64.6 | 83.3 | 90.4 | 86.5 | 80.2 | 75.5 | 80.3 | 86.0 | 87.1 | 81.4 | 59.6 | 79.5 |
| | | GaitGP(ours) | **70.0** | **83.7** | **91.3** | **89.3** | **81.6** | **77.4** | **82.0** | **88.8** | **91.4** | **86.0** | **67.5** | **82.6** |
| | BG#1-2 | GaitSet [3] | 55.8 | 70.5 | 76.9 | 75.5 | 69.7 | 63.4 | 68.0 | 75.8 | 76.2 | 70.7 | 52.5 | 68.6 |
| | | GaitGP(ours) | **62.1** | **74.2** | **79.0** | **76.5** | **72.5** | **64.1** | **69.7** | **77.0** | **79.2** | **74.3** | **58.2** | **71.5** |
| | CL#1-2 | GaitSet [3] | 29.4 | 43.1 | 49.5 | **48.7** | 42.3 | 40.3 | **44.9** | **47.4** | **43.0** | **35.7** | 25.6 | 40.9 |
| | | GaitGP(ours) | **31.8** | **47.3** | **50.9** | 48.1 | **46.2** | **41.8** | 44.4 | 44.3 | 42.9 | 34.8 | **28.3** | **41.9** |
| MT(62) | NM#5-6 | GaitNet [46] | 49.3 | 61.5 | 64.4 | 63.6 | 63.7 | 58.1 | 59.9 | 66.5 | 64.8 | 56.9 | 44.0 | 59.3 |
| | | MGAN [25] | 54.9 | 65.9 | 72.1 | 74.8 | 71.1 | 65.7 | 70.0 | 75.6 | 76.2 | 68.6 | 53.8 | 68.1 |
| | | GaitSet [3] | 86.8 | 95.2 | 98.0 | 94.5 | **91.5** | 89.1 | 91.1 | 95.0 | 97.4 | 93.7 | 80.2 | 92.0 |
| | | GaitGP(ours) | **88.9** | **95.3** | **98.2** | **97.4** | 91.4 | **90.0** | **92.7** | **98.4** | **98.7** | **94.3** | **85.2** | **93.7** |
| | BG#1-2 | GaitNet [46] | 29.8 | 37.7 | 39.2 | 40.5 | 43.8 | 37.5 | 43.0 | 42.7 | 36.3 | 30.6 | 28.5 | 37.2 |
| | | MGAN [25] | 48.5 | 58.5 | 59.7 | 58.0 | 53.7 | 49.8 | 54.0 | 51.3 | 59.5 | 55.9 | 43.1 | 54.7 |
| | | GaitSet [3] | 79.9 | 89.8 | 91.2 | 86.7 | 81.6 | 76.7 | 81.0 | 88.2 | 90.3 | 88.5 | 73.0 | 84.3 |
| | | GaitGP(ours) | **80.9** | **87.7** | **91.9** | **90.6** | **85.1** | **77.9** | **81.9** | **90.5** | **94.5** | **89.5** | **77.7** | **86.2** |
| | CL#1-2 | GaitNet [46] | 18.7 | 21.0 | 25.0 | 25.1 | 25.0 | 26.3 | 28.7 | 30.0 | 23.6 | 23.4 | 19.0 | 24.2 |
| | | MGAN [25] | 23.1 | 34.5 | 36.3 | 33.3 | 32.9 | 32.7 | 34.2 | 37.6 | 33.7 | 26.7 | 21.0 | 31.5 |
| | | GaitGP(ours) | **54.8** | **63.5** | **72.4** | **67.4** | **61.7** | **58.7** | **60.8** | **64.9** | **64.6** | **59.8** | **48.8** | **61.6** |
| LT(74) | NM#5-6 | CNN-LB [20] | 82.6 | 90.3 | 96.1 | 94.3 | 90.1 | 87.4 | 89.9 | 94.0 | 94.7 | 91.3 | 78.5 | 89.9 |
| | | GaitNet [46] | 91.2 | 92.0 | 90.5 | 95.6 | 86.9 | 92.6 | 93.5 | 96.0 | 90.9 | 88.8 | 89.0 | 91.6 |
| | | GaitSet [3] | 90.8 | 97.8 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
| | | ACL [33] | **92.0** | **98.5** | **100.0** | **98.9** | **95.7** | 91.5 | 94.5 | 97.7 | 98.4 | 96.7 | **91.9** | 96.0 |
| | | GaitGP(ours) | 91.7 | 98.2 | 98.8 | 98.3 | **95.7** | **93.4** | **95.9** | **99.4** | **99.1** | **98.3** | 89.6 | **96.2** |
| | BG#1-2 | CNN-LB [20] | 64.2 | 80.6 | 82.7 | 76.9 | 64.7 | 63.1 | 68.0 | 76.9 | 82.2 | 75.4 | 61.3 | 72.4 |
| | | GaitNet [46] | 83.0 | 87.8 | 88.3 | **93.3** | 82.6 | 74.8 | **89.5** | 91.0 | 86.1 | 81.2 | **85.6** | 85.7 |
| | | GaitSet [3] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
| | | GaitGP(ours) | **87.1** | **91.9** | **94.6** | 92.2 | **88.9** | **82.7** | 86.2 | **94.1** | **96.5** | **94.5** | 84.7 | **90.3** |
| | CL#1-2 | CNN-LB [20] | 37.7 | 57.2 | 66.6 | 61.1 | 55.2 | 54.6 | 55.2 | 59.1 | 58.9 | 48.8 | 39.4 | 54.0 |
| | | GaitNet [46] | 42.1 | 58.2 | 65.1 | 70.7 | 68.0 | 70.6 | 65.3 | 69.4 | 51.5 | 50.1 | 36.6 | 58.9 |
| | | GaitGP(ours) | **60.0** | **74.0** | **78.6** | **77.8** | **69.5** | **67.7** | **69.7** | **72.3** | **73.5** | **66.6** | **51.5** | **69.2** |

*Note:* The parts in bold are the best experimental results.

Abbreviations: BG, bag; CASIA-B, the CASIA gait recognition dataset B; CL, cloth; LT, large-scale training; MT, medium-scale training; NM, normal; ST, small-scale training.

$64 \times 64$. The optimizer Adam is Adopted [47] to perform gradient optimization and the learning rate is set to $1e-4$. In addition, the momentum is set to 0.9 and the margin of the Separate Triplet Loss is set to 0.2, the same as [44]. In CASIA-B, we set the batch size to $(8, 16)$, and the number of training iterations is $90K$. In OU-MVLP, because it contains far more sequences than CASIA-B, we set the number of parts of the GPConv layer in block2 and block3 to 2, 2, 4, 4, and the batch size to $(32, 8)$, the number of iterations is set to $250K$, and the learning rate is set to 1*e*-5.

## 4.2 | Comparison with the state-of-art methods

### 4.2.1 | CASIA-B

As shown in Table 3, we compare our method with the latest gait recognition methods, which mainly include CNN-LB [20], GaitNet [46], GaitSet [3], MGAN [25] and ACL [33]. To make a systematic and comprehensive comparison with the advanced methods, all conditions (NM, BG, CL) are included, and

further experiments and comparative analyses are carried out with different training sample sizes (BT, MT, LT). The proposed method achieves the best recognition accuracy in almost all angles.

(1)As shown in Table 3, CNN-LB [20] is a GEI-based method and others are all based on the silhouettes, but the latter all perform better than the former. It shows that video-based methods have great potential in extracting more fine-grained information and distinguishing information from images.
(2)We discuss with GaitNet [46] and MGAN [25], which have the same structural purpose but different architecture composition. In GaitNet [45], the Auto-Encoder is introduced to obtain more distinguishing functions, and the multi-layer LSTM is applied for spatio-temporal modeling. MGAN uses a generative confrontation network to map different costumes to the same template from the front and side perspectives. In our model, we introduce the CSP to extract the local feature attention through channel-level division as the spatio-temporal attention feature of the subject.
(3)Compared with GaitSet [3], our structure is used a partitionable convolution unit called the PConv, which is used to obtain the channel-level feature fusing with spatial attention. The MGM of GaitGP also has a similar structure as that of GaitSet, but the MGM pays more attention to the fine-grained local segmentation using the independent operation to enlarge more representative features and reduce the similarity between different subjects. This result reveals the advantages of the PConv and MGM through experiments. From the experiment, GaitGP has obtained better results under various walking conditions on CASIA-B.

### 4.2.2 | OU-MVLP

To prove the effectiveness of our method, we conduct two large-scale experiments on OUMVLP. (1) We use the same evaluation setting as [12] where 5153 people are trained and 5154 people are tested. The silhouettes of four typical views (0°, 30°, 60°, 90°) are evaluated for cross-view recognition, as shown in Table 4. (2) We list the results in two gallery collections including all the 14 views and the results are averaged on the gallery view (exclude the identical-view). We set the dimension of the global feature and the local feature as 512 and reduce the dimensionality through MGM to 256, as shown in Table 5.

### 4.2.3 | OULP

To prove the broad applicability of our method, we also perform the experiments on OULP. The results are shown in Table 6: We compare our method with CNN-LB [20], GEINet [46], and MGAN [25]. These methods are based on cross-view to calculate the accuracy, that is, calculating the average accuracy of each view angle excluding the same view angle. Our GaitGP performs better than these methods.

**TABLE 4** OUMVLP results excluding the identical-view cases under four typical views (0°, 30°, 60°, 90°)

| | Gallery 0°, 30°, 60°, 90° | | |
|---|---|---|---|
| **Probe** | **GEINet [46]** | **3in + 2diff [12]** | **GaitGP(ours)** |
| 0° | 8.2 | 25.5 | **73.8** |
| 30° | 32.3 | 50.0 | **87.3** |
| 60° | 33.6 | 45.3 | **84.5** |
| 90° | 28.5 | 40.6 | **83.9** |
| Mean | 25.7 | 40.4 | **83.4** |

*Note:* The parts in bold are the best experimental results.

**TABLE 5** OUMVLP results excluding the identical-view cases under all views

| | Gallery all 14 views | |
|---|---|---|
| **Probe** | **GEINet [46]** | **GaitGP(ours)** |
| 0° | 11.4 | **73.8** |
| 15° | 29.1 | **84.0** |
| 30° | 41.5 | **87.3** |
| 45° | 45.5 | **87.6** |
| 60° | 39.5 | **84.5** |
| 75° | 41.8 | **86.2** |
| 90° | 38.9 | **83.9** |
| 180° | 14.9 | **77.4** |
| 195° | 33.1 | **82.5** |
| 210° | 43.2 | **85.9** |
| 225° | 45.6 | **86.3** |
| 240° | 39.4 | **82.5** |
| 255° | 40.5 | **84.6** |
| 270° | 36.3 | **81.6** |
| Mean | 35.8 | **83.4** |

*Note:* The parts in bold are the best experimental results.

## 4.3 | Ablation study

To further verify the effectiveness of each component in our proposed network GaitGP, the two components of the CAFE, PConv and CSP, and the MGM module in the GPFC pipeline are included. We perform the ablation study of these components on the CASIA-B data set. Research, experimental results, and analysis are as follows.

**TABLE 6** OULP cross-view average accuracies (%) for all pairs of four view angles

| Probe angle | Method | Gallery angle | | | | | |
|---|---|---|---|---|---|---|---|
| | | 55° | 65° | 75° | 85° | Mean | Identical angle |
| 55° | CNN-LB [20] | - | 98.3 | 96.0 | 80.5 | 91.6 | 98.8 |
| | GEINet [46] | - | 93.2 | 89.7 | 79.9 | 87.6 | 94.7 |
| | MGAN [25] | - | **99.4** | 96.1 | 77.9 | 91.6 | 98.8 |
| | Method [13] | - | - | - | - | - | 84.7 |
| | GaitGP(ours) | - | 98.9 | **96.8** | **90.3** | **95.3** | **99.6** |
| 65° | CNN-LB [20] | 96.3 | - | 97.3 | 83.3 | 92.3 | 98.9 |
| | GEINet [46] | 93.7 | - | 93.8 | 90.6 | 92.7 | 95.1 |
| | MGAN [25] | 97.7 | - | **98.5** | 84.4 | - | - |
| | Method [13] | - | - | - | - | - | 86.6 |
| | GaitGP(ours) | **99.0** | - | 97.1 | **90.8** | **95.6** | **99.1** |
| 75° | CNN-LB [20] | 94.2 | 97.8 | - | 85.1 | 92.4 | 98.9 |
| | GEINet [46] | 90.1 | 94.1 | - | **93.8** | 92.7 | 97.7 |
| | MGAN [25] | 94.8 | 97.8 | - | 86.4 | - | - |
| | Method [13] | - | - | - | - | - | 86.9 |
| | GaitGP(ours) | **98.6** | **98.7** | - | 91.1 | **96.1** | **99.1** |
| 85° | CNN-LB [20] | 90.0 | 96.0 | 98.4 | - | 94.8 | 98.9 |
| | GEINet [46] | 81.4 | 91.2 | 94.6 | - | 89.1 | 94.7 |
| | MGAN [25] | 86.9 | 97.4 | **99.5** | - | - | - |
| | Method [13] | - | - | - | - | - | 85.7 |
| | GaitGP(ours) | **97.5** | **98.0** | 97.1 | - | **97.5** | **99.2** |

*Note:* The parts in bold are the best experimental results.

### 4.3.1 | Effectiveness of PConv

As introduced in Table 1, we present the parameter settings of the PConv. To evaluate the robustness, we design four groups of experiments. The blocks in the CAFE are composed of two PConv. In Exp.1-1, we set the $\tau = 1$ in Block1 for retaining its original state and the remaining two blocks are parameterized as 1. The difference between Exp.1-2 and Exp.1-1 is that Block1 remains unchanged, but the parameter $\tau$ in Block2 and Block3 are set to 2 and 4. Exp.1-3 is based on Exp.1-2, but the latter two blocks are sets to 4. Similarly, the parameter $\tau$ of Block2 and Block3 in Exp.1-4 are set to 4 and 8. All the results of these controlled experiments are shown in Table 7.

Comparing Exp.1-1 and Exp.1-2, on the one hand, we found that the blocks with the original state($\tau = 1$) are not effective, which shows the advantages of partitionable extraction. On the other hand, the features of Exp.1-4 are too dispersed and lead to poor performance in the superficial layer, which is probably because of too much subdivision destroying the information of silhouette between the edges of the adjacent regions and increases the proportion of noise covariates. Finally, by comparing the differences of Exp.1-2, Exp.1-3, Exp.1-4, we observe that the average rank-1 accuracy first rises and then falls on the NM and BG subset, while it continues to rise under the CL subsets. It is believed that the reason for this phenomenon is that the different receptive fields of the top neurons can adapt to different walking conditions.

### 4.3.2 | Effectiveness of CSP

The traditional spatial feature mapping [3, 4] uses $Max(\cdot)$ or $Avg(\cdot)$ to aggregate spatial information. But, using them alone cannot realize the mapping adaptively. In this paper, we introduce CSP to achieve spatial feature mapping. Figure 3 shows its internal structure and describes the components used inside. Inspired by the idea in [1], slicing the feature map at the channel level, we design a new statistical function *SP* and use *ConvNet* to weight the local features to enhance attention. To verify the effectiveness of CSP, we design comparative experiments by implementing methods with different spatial feature mapping strategies on the CASIA-B data set. Note that the channel-level slice parameters are referred to the parameters $\tau$ in the previous ablation experiment.

The results are shown in Table 8. Exp.2-1 uses the traditional statistical function *SP* under the conditions of NM and BG. Compared with Exp.2-2, we set the parameter $\tau$ of $SP_1$, $SP_2$, $SP_3$ to slice 1, 2, and 4 in different blocks, which has better performance. In Exp.2-3, the addition of the *ConvNet* layer aims to enhance the attention and make the aggregation of spatial

**TABLE 7** The ablation experiment performed on CASIA-B using the setup LT. The result is the average level 1 accuracy of all 11 views, excluding the case of the same view. Comparison of different parameter settings of PConv

| | $\tau$ of PConv | | | | | |
|---|---|---|---|---|---|---|
| **Exp1** | $\tau_{Block_1}$ | $\tau_{Block_2}$ | $\tau_{Block_3}$ | **NM** | **BG** | **CL** |
| 1-1 | 1 | 1 | 1 | 95.3 | 88.8 | 68.1 |
| 1-2 | 1 | 2 | 4 | **96.2** | **90.3** | 67.8 |
| 1-3 | 1 | 4 | 4 | 96.0 | 89.8 | 68.7 |
| 1-4 | 1 | 4 | 8 | 95.9 | 90.2 | **69.2** |

*Note:* The parts in bold are the best experimental results.

Abbreviation: BG, bag; CASIA-B, the CASIA gait recognition dataset B; CL, cloth; LT, large-scale training; NM, normal.

**TABLE 8** The ablation experiment performed on CASIA-B using setup LT. Results are rank-1 accuracies of all 11 views, excluding the case of the same view. Comparison of CSP with different settings for different blocks

| | $\tau$ of CSP | | | | | | |
|---|---|---|---|---|---|---|---|
| **Exp2** | $\tau_{SP_1}$ | $\tau_{SP_2}$ | $\tau_{SP_3}$ | **ConvNet** | **NM** | **BG** | **CL** |
| 2-1 | 1 | 1 | 1 | | 93.4 | 82.8 | 59.1 |
| 2-2 | 1 | 2 | 4 | | 94.3 | 86.5 | 64.5 |
| 2-3 | 1 | 2 | 4 | ✓ | **96.2** | **90.3** | 67.8 |
| 2-4 | 1 | 4 | 8 | ✓ | 95.9 | 90.2 | **69.2** |

*Note:* The parts in bold are the best experimental results.

Abbreviations: BG, bag; CASIA-B, the CASIA gait recognition dataset B; CL, cloth; CSP, Channel-level Spatial Pooling; LT, large-scale training.

information more effective, reaching accuracy rates of 96.2% and 90.3%. Besides, in the CL setting, when the parameter $\tau$ of $SP_1$, $SP_2$, $SP_3$ is set to 1, 4, and 8, as shown in Exp.2-4, the highest accuracy rate is 69.2%. This may indicate that fine-grained feature extraction is better for extracting silhouette maps with bags.

### 4.3.3 | Effectiveness of the MGM

We duplicate five branches of the intermediate feature maps obtained by the backbone network, named $Part_1$, $Part_2$, $Part_3$, $Part_4$ and $Part_5$, and the corresponding configurations are shown in Table 2. From the experimental results, we found that setting the horizontal stripes as $\rho = 2^{m-1}$, $m \in 1, 2, 3, 4, 5$, the same as in [3], performs well, which shows that different fine-grained segmentation can better capture details that are easily ignored for recognition.

In our experiment, we explore the influence of multi-branch architecture from two aspects. As shown in Table 9, on the one hand, the structure with only one partition branch $Part_1$(considered as the global representation) is compared with the structure of integrating only four independent different multi-granularity branches. It is shown that the integrated strategy achieves better performance than any single participating network. It shows that, compared with the global network, the collaborative learning of branches has more discriminatory feature representations. On the other hand, we combine the two structures and compare them with the above two experiments. The effect of combining the global features and local features is higher than using one of them alone. We believe that the mutual influence between the four independent different multi-granularity branches supplements their blind spots in their learning process.

**TABLE 9** The ablation experiment performed on CASIA-B using setup LT. The result is the average level 1 accuracy of all 11 views, excluding the case of the same view. Accuracy (%) of using different branches in MGM

| $Part_1$ | $Part_2$ | $Part_3$ | $Part_4$ | $Part_5$ | **NM** | **BG** | **CL** |
|---|---|---|---|---|---|---|---|
| ✓ | | | | | 80.4 | 70.3 | 43.2 |
| | ✓ | ✓ | ✓ | ✓ | 90.6 | 80.4 | 57.8 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **96.2** | **90.3** | **69.2** |

*Note:* The parts in bold are the best experimental results.

Abbreviation: BG, bag; CASIA-B, the CASIA gait recognition dataset B; CL, cloth; LT, large-scale training; MGM, Multi-Granularity Mapping; NM, normal.

**TABLE 10** The ablation experiment performed on CASIA-B using setup LT. The result is the average level 1 accuracy of all 11 views, excluding the case of the same view. Accuracy (%) of using different branches in MGM

| **Method** | **CASIA-B** |
|---|---|
| Wu et al.[49] | 40 min |
| Zhang et al.[48] | 3 min |
| Zhang et al.[32] | 1.5 min |
| GaitGP(ours) | **1.36** min |

*Note:* The parts in bold are the best experimental results.

Abbreviation: CASIA-B, the CASIA gait recognition dataset B; LT, large-scale training.

### 4.3.4 | Efficiency of GaitGP

As discussed in [48], the efficiency of the pair-wise simulation degree learning method [49] is limited. On the other hand, since each sample only needs to be calculated once [3], our network takes 1.36 min to complete the test on 4 NVIDIA 1080TI GPUs. Table 10 lists the efficiency comparison on CASIA-B.

## 5 | CONCLUSION

This paper proposes a new network architecture and designs the PConv to extract the global and partial features by combining the advantages of both. We also propose CSP for spatial learning attention and feature expression to improve the performance of gait recognition tasks. In addition, through the multi-granularity horizontal segmentation pipeline, MGM, different multi-granularity branches are integrated to obtain the final gait representation. Experimental

results on three public datasets verify the effectiveness and efficiency of our method.

## ACKNOWLEDGEMENTS

This work was partially supported by the Natural Science Foundation of Guangdong Province No. 2018A030313318 and the Key-Area Research and Development Program of Guangdong Province No. 2019B111101001.

## ORCID

*Jing Xiao* https://orcid.org/0000-0002-5242-7909

## REFERENCES

1. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2285–2294 (2018)
2. Zhang, K., et al.: Learning joint gait representation via quintuplet loss minimization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4700–4709 (2019)
3. Chao, H., et al.: GaitSet: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8126–8133. AAAI (2019)
4. Fan, C., et al.: Gaitpart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225–14233 (2020)
5. Sun, Y., et al.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 480–496 (2018)
6. Li, S., et al.: Diversity regularized spatiotemporal attention for video-based person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 369–378 (2018)
7. Luo, H., et al.: Bag of tricks and a strong baseline for deep person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (2019)
8. Fu, Y., et al.: Horizontal pyramid matching for person re-identification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8295–8302. AAAI (2019)
9. Yao, H., et al.: Deep representation learning with part loss for person re-identification. IEEE Trans. Image Process. 28, 2860–2871 (2019)
10. Lin, B., et al.: Learning effective representations from global and local features for cross-view gait recognition. arXiv preprint arXiv:2011.01461 (2020)
11. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), vol. 4, pp. 441–444. IEEE (2006)
12. Takemura, N., et al.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans. Comput. Vis. Appl. 10, 4 (2018)
13. Iwama, H., et al.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. IEEE Trans. Inf. Forensics Secur. 7, 1511–1521 (2012)
14. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3d convolutional neural networks. In: Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), pp. 4165–4169. IEEE (2016)
15. Zhao, G., et al.: 3d gait recognition using multiple cameras. In: Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pp. 529–534. IEEE (2006)
16. Ariyanto, G., Nixon, M.S.: Model-based 3d gait biometrics. In: Proceedings of the 2011 International Joint Conference on Biometrics (IJCB), pp. 1–7. IEEE (2011)
17. Feng, Y., Li, Y., Luo, J.: Learning effective gait features using lstm. In: Proceedings of the 2016 23rd International Conference on Pattern Recognition (ICPR), pp. 325–330. IEEE (2016)
18. Zhang, J., et al.: Low-resolution gait recognition. IEEE Trans. Syst. Man Cybern., Part B (Cybern.). 40, 986–996 (2010)
19. Hu, M., et al.: View-invariant discriminative projection for multi-view gait-based human identification. IEEE Trans. Inf. Forensics Secur. 8, 2034–2045 (2013)
20. Wu, Z., et al.: A comprehensive study on cross-view gait based human identification with deep cnns. IEEE Trans. Pattern Anal. Mach. Intell. 39, 209–226 (2016)
21. Takemura, N., et al.: On input/output architectures for convolutional neural network-based cross-view gait recognition. IEEE Trans. Circ. Syst. Video Technol. 29, 2708–2719 (2017)
22. Yu, S., et al.: Invariant feature extraction for gait recognition using only one uniform model. Neurocomputing. 239, 81–93 (2017)
23. Liao, R., et al.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: Proceedings of the Chinese Conference on Biometric Recognition, pp. 474–483. Springer (2017)
24. Connor, P., Ross, A.: Biometric recognition by gait: a survey of modalities and features. Comput. Vis. Image Understand. 167, 1–27 (2018)
25. He, Y., et al.: Multi-task gans for view-specific feature learning in gait recognition. IEEE Trans. Inf. Forensics Secur. 14, 102–113 (2018)
26. Goodfellow, I.J., et al.: Generative adversarial networks. arXiv preprint arXiv:1406.2661 (2014)
27. Yu, S., et al.: Gaitgan: invariant gait feature extraction using generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 30–37 (2017)
28. Rokanujjaman, M., Hossain, M.A., Islam, M.R.: Effective part selection for part-based gait identification. In: Proceedings of the 2012 7th International Conference on Electrical and Computer Engineering, pp. 17–19. IEEE (2012)
29. Rida, I., Jiang, X., Marcialis, G.L.: Human body part selection by group lasso of motion for model-free gait recognition. IEEE Signal Process. Lett. 23, 154–158 (2015)
30. Su, C., et al.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3960–3969 (2017)
31. Wang, G., et al.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia, pp. 274–282 (2018)
32. Zhang, Y., et al.: Cross-view gait recognition by discriminative feature learning. IEEE Trans. Image Process. 29, 1001–1015 (2019)
33. Zhang, Z., et al.: Gait recognition via disentangled representation learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4710–4719 (2019)
34. Li, D., et al.: Learning deep context-aware features over body and latent parts for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 384–393 (2017)
35. Cheng, D., et al.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1335–1344 (2016)
36. Jaderberg, M., et al.: Spatial transformer networks. arXiv preprint arXiv:1506.02025 (2015)
37. Liu, X., et al.: Hydraplus-net: attentive deep features for pedestrian analysis. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 350–359 (2017)
38. Yang, Q., et al.: Patch-based discriminative feature learning for unsupervised person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3633–3642 (2019)
39. Zhao, L., et al.: Deeply-learned part-aligned representations for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3219–3228 (2017)
40. Wang, X., et al.: Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018)

41. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the International Conference on Machine Learning, pp. 2048–2057. PMLR (2015)
42. Cao, Y., et al.: Gcnet: non-local networks meet squeeze-excitation networks and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (2019)
43. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141 (2018)
44. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
45. Li, W., Zhu, X., Gong, S.: Person re-identification by deep joint learning of multi-loss classification. arXiv preprint arXiv:1705.04724 (2017)
46. Shiraga, K., et al.: Geinet: view-invariant gait recognition using a convolutional neural network. In: Proceedings of the 2016 International Conference on Biometrics (ICB), pp. 1–8. IEEE (2016)
47. Kingma, D.P., Ba, J., Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
48. Zhang, Y., et al.: A comprehensive study on gait biometrics using a joint cnn-based method. Pattern Recogn. 93, 228–236 (2019)
49. Wu, Z., Huang, Y., Wang, L.: Learning representative deep features for image set analysis. IEEE Trans. Multimed. 17, 1960–1968 (2015)

**How to cite this article:** Xiao, J., et al.: Learning discriminative representation with global and fine-grained features for cross-view gait recognition. CAAI Trans. Intell. Technol. 7(2), 187–199 (2022). https://doi.org/10.1049/cit2.12051