TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 25/31 pp345–356
DOI: 10.26599/TST.2023.9010089
Volume 30, Number 1, February 2025

GaitFFDA: Feature Fusion and Dual Attention Gait Recognition Model

Zhixiong Wu* and Yong Cui

Abstract: Gait recognition has a wide range of application scenarios in the fields of intelligent security and transportation. Gait recognition currently faces challenges: inadequate feature methods for environmental interferences and insufficient local-global information correlation. To address these issues, we propose a gait recognition model based on feature fusion and dual attention. Our model utilizes the ResNet architecture as the backbone network for fundamental gait features extraction. Subsequently, the features from different network layers are passed through the feature pyramid for feature fusion, so that multi-scale local information can be fused into global information, providing a more complete feature representation. The dual attention module enhances the fused features in multiple dimensions, enabling the model to capture information from different semantics and scale information. Our model proves effective and competitive results on CASIA-B (NM: 95.6%, BG: 90.9%, CL: 73.7%) and OU-MVLP (88.1%). The results of related ablation experiments show that the model design is effective and has strong competitiveness.

Key words: gait recognition; neural network; attention mechanism

1 Introduction

Gait is a biometric characteristic closely associated with individuals. It reflects unique features of individuals during their walking process, including both inherent traits such as limb lengths and proportions, and walking habits like the amplitude and frequency of limb Swings. Compared to other biometric features like fingerprints, voiceprints, and faces, gait possesses stronger anti-counterfeiting capabilities, encompassing both innate and acquired charac-teristics. It is more convenient to sample as it does not require participants' cooperation, and it has lower data quality requirements, with lower data resolution needed for collection, and is not restricted by time and space.

• Zhixiong Wu and Yong Cui are with Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China. E-mail: wzx@linewell.com.
* To whom correspondence should be addressed.
Manuscript received: 2023-08-04; revised: 2023-08-22; accepted: 2023-08-26

As a result, gait analysis has found applications in diverse fields, including biometric recognition, intelligent security, medical rehabilitation, and behavior analysis. Its potential for application in a broader range of scenarios has been proven through related research.

A common preprocessing method for gait data is to binarize the video images or segment the foreground and background to obtain silhouette data of the human body. This silhouette data is then used for subsequent feature extraction. After constructing gait data, further processing of its spatio-temporal features is necessary. Effectively mining the spatio-temporal features of gait data and exploring efficient representations of gait features will be the key foundation for subsequent modeling.

Various methods that can be divided into two categories have been proposed to deal with the obstacles. The first category of methods[1–4] aggregating silhouette images within one gait cycle into a single template image. For example, GEI (Gait

Energy Image)[5] is a gait energy map generated by normalizing and averaging sequential data.

The second category of methods[6−10] involves feeding a set of silhouette images within one gait cycle as input data into a subsequent network for feature modeling. Such input data can be either ordered sequences or unordered sets.

The template-based feature representation method focuses on summarizing gait features from a global perspective. Its advantage lies in its concise expression and ease of operation for the network. However, its disadvantage is that it may overlook gait feature details, making it less effective for extracting compact feature representations.

On the other hand, the sequence-based feature representation method emphasizes extracting feature representations from a local perspective, capturing features from single frames or body parts, and establishing inter-frame correlations. However, this approach tends to be more complex in its processing.

Therefore, the existing gait recognition methods have two main problems. One is that the feature extraction methods are not adequate enough to address various environmental interferences, which means the gait details are often ignored, and further feature processing is lacking. The second is the trade-off between local and global information, which requires necessity to balance weighting and establish effective correlation between the two.

To address these problems, we propose a Gait recognition model based on Feature Fusion and Dual Attention (GaitFFDA). Using a set of unordered gait silhouettes, a ResNet architecture backbone is utilized to construct the fine-grained features of gait. The output feature maps of layer blocks are fed into a pyramid pooling module that combines dual attention mechanism to learn feature representations more efficiently while preserving local and global information. Specifically, our work includes the following aspects:

● To address the issue of inadequate feature extraction methods, we adopt a ResNet architecture backbone to extract frame-level features of gait and using a temporal pooling layer to construct set-level features.

● We design and employ a feature fusion module for our model, so that multi-scale local information can be fused into global information, and provide a more complete feature representation for network learning. This module can effectively integrate local and global information.

● To further ensure the propagation and utilization of local and global information, a dual attention module is designed, which consists of dimension attention and scale attention. Dimension attention is used to capture and learn information from different semantic and spatial dimensions, while scale attention is used to selectively establish representation relationships between gait identity and various scale information.

● The ablation experiments conducted on two datasets, CASIA-B and OU-MVLP, demonstrated the efficacy of our network design. And comparing the performance with other models, our model has strong competitiveness.

2 Related Work

In recent years, gait recognition has gained significant popularity as a research topic. Researchers have introduced a plethora of innovative ideas and methods to construct gait data and develop efficient feature representations.

In general, gait recognition data is obtained from videos or images recorded by cameras. Depending on the different data construction methods, the primary approaches include skeleton-based methods and silhouette-based methods for classification.

The skeleton-based method[11−14] utilizes pose recognition frameworks to locate the spatial positions of human skeletal keypoints. These keypoints are then connected to form a spatial and temporal graph network structure data. Based on the constructed structural data, further data preprocessing is performed.

Liu et al.[112] extracted distance and angle information between keypoints as spatio-temporal features. Similarly, Ahmed et al.[14] used relative angles between the most relevant joints in a gait sequence as input features for their model.

Although the skeleton-based method can accurately capture joint changes during human walking and eliminate the influence of irrelevant attributes such as clothing and appearance, this approach requires significant computational resources and heavily relies on the recognition performance of the pose recognition framework.

The silhouette-based method involves binarizing video images or segmenting foreground and

background to obtain silhouette data of the human body, as shown in Fig. 1. In their work, Fan et al.[8] divided the silhouette images into four equally sized parts and modeled the spatio- temporal features of each part separately. On the other hand, Feng et al.[15] utilized convolutional networks to generate heatmaps for 12 different body parts of the human silhouette.

Compared to the skeleton-based method, the silhouette-based method also eliminates the influence of appearance attributes and possesses the capability to extract information such as speed, frequency, and step length. Additionally, it requires fewer computational resources for data generation. Therefore, in related research, most studies have adopted the silhouette-based method as the initial processing step for video images.

Regarding gait data constructed based on silhouette, the common feature representation methods are template representation and sequence representation. Template representation[16−19] refers to the method of aggregating silhouette images within one gait cycle into a single template image.

In Wang et al.'s research[16], they utilized color mapping functions to encode the gait silhouette sequence and fused it into a unified Chrono-Gait Image (CGI) template. Ghaeminia and Shokouhi[17] introduced the concept of encoding gait's motion energy as a Gait Salient Image (GSI) template. Motion features were extracted using suitable spatio-temporal filters and then averaged over the gait cycle.

However, using only a single template image makes it difficult to effectively retain the temporal and contextual information between frames, and some action details might be lost. To address this challenge, other scholars[6, 10, 20−22] have proposed methods based on sequence representation.

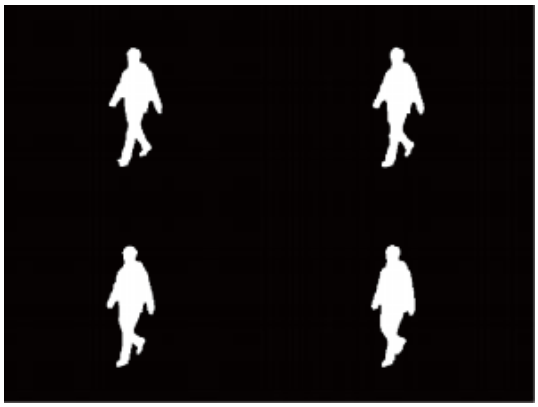

Fig. 1 2D gait silhouettes.

Lin et al.[20] developed a gait recognition framework that incorporates multiple temporal scales to capture temporal information effectively. They utilized a 3D Convolutional Neural Network (CNN) network to extract spatial-temporal features at various time scales, effectively combining frame and interval fusion information. However, the 3D CNN network requires fixed-scale gait sequences as network input, which limits its flexibility. Chao et al.[4] proposed a method that takes gait data as an unordered set input and utilizes a CNN network to extract frame-level features. Subsequently, these features are aggregated using set pooling operations. Although this approach can bypass sequence length and order constraints, the dimension of the aggregated features becomes too high for the model to effectively distill knowledge.

Template feature methods encode the whole period of gait sequence data, thus the global information are more complete than the other methods. While frame-level feature methods operate on each gait image sepearately, they could generate richer local information. Whether the method is based on template feature or frame-level feature, the model design needs to consider the tradeoff between local information and global information. To tackle this problem, a reasonable and practical solution would be adopting an effective network design to combine the advantage of two methods.

The appropriate feature representation method involves the preliminary induction and summarization of gait features, and it determines how subsequent models will be constructed. Template representation places emphasis on summarizing gait features from a global perspective, which offers the advantage of a concise expression and ease of operation for the network. On the other hand, sequence representation focuses on extracting feature representations from a local perspective, considering single frames or body parts, and establishing correlations between frames.

3 Method

3.1 Model structure

Based on the aforementioned analysis, this study proposes a gait recognition model based on feature fusion and dual attention.

The model structure is shown in Fig. 2. The model takes unordered sets of silhouette data as input. The

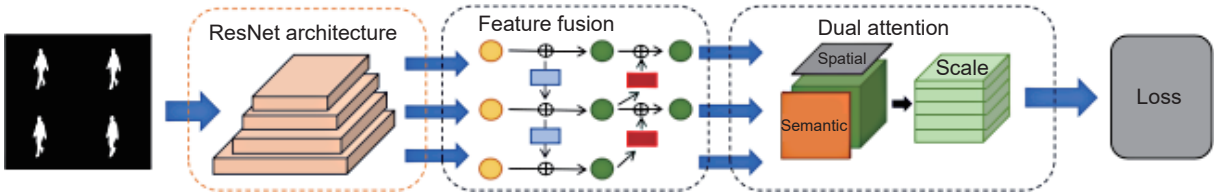

Fig. 2 Overall structure of proposed model GaitFFDA. In the feature fusion module, the yellow circles represent 1×1 convolution operations, the blue rectangles represent upsampling operations, the green circles represent 3×3 convolution operations, and the red rect-angles represent downsampling operations. In the dual attention module, the feature map is sequentially analyzed for dimension attention and spatial attention. The dimension attention primarily focuses on capturing the attention across the semantic dimensions and spatial dimensions of the feature map.

silhouette images in the set are partitioned at the frame level, and all the images contained in a set (usually representing one gait cycle) are simultaneously fed into a ResNet architecture network. The convolutional layers of the ResNet network are used to extract feature information from individual silhouette frames, forming preliminary frame feature maps. These feature maps are then forwarded in two directions: firstly, from one ResNet block to the next ResNet block, and secondly, the feature maps output from each network block are sent to the feature fusion module.

In the feature fusion module, the output feature maps at different layers are fused at different scales in top-down and bottom-up ways to obtain compact global information representation.

The design of dual attention module unfolds based on the sequence of data propagation. Firstly, to focus on local gait features, the model incorporates dimension attention, enhancing gait features at different semantic and spatial dimensions. As the mapping dimension is relatively large at this stage, a scale attention mechanism is introduced to improve learning efficiency, enabling the model to selectively learn from the mapping results.

Subsequently, the model is trained using a combination loss, which includes both cross-entropy loss and triplet loss. Further details on the network design and feature propagation process will be introduced in the following chapters.

3.2 ResNet architecture network

To extract fundamental gait features, we follow the work of Fan et al.[23] by implementing a ResNet architecture backbone in the network.

Specifically, the ResNet-based network includes a shallow convolutional layer and four convolutional blocks. Each convolutional block consists of two convolutional layers, and each layer employs a kernel size of 3 × 3. The output feature dimensions of each block are 64, 128, 256, and 512, respectively. The feature maps generated by each convolutional block propagate in two directions. As shown in Fig. 3, the first direction involves processing the frame-level features through two layers of max-pooling convolutional layers in both the 2D spatial dimension and the temporal dimension, resulting in set-level features, which are then passed into the feature fusion module. The second direction is to directly forward the feature maps to the next convolutional block.

As the network deepens and evolves through different stages, there are two aspects of changes in information extraction. Firstly, there is a variation in information hierarchy. Information extracted in the shallow convolutional blocks pertains to lower-level features, closely related to gait contour and body posture representation. As the network transitions from

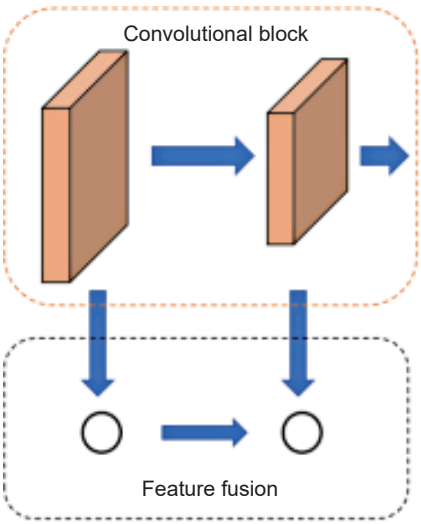

Fig. 3 Illustration of feature maps propagation direction. Horizontally, the feature maps are propagated from the preceding convolutional block to the subsequent one. Vertically, the feature maps of each convolutional block are passed into the corresponding feature fusion module level.

lower-level to higher-lever information, the deep convolutional blocks extract higher-level features, which are more closely related to the identity of the gait subject.

Secondly, there is a change in information dimensionality, with feature map sizes and channel numbers increasing step by step, providing the network with the ability to represent complex information. However, this also requires the model to possess the capability to discern relevant and effective information. Effectively utilizing the representations from different stages is crucial for the model to decouple identity information.

3.3 Feature fusion module

As described earlier, the feature output of each convolutional block is not only propagated to the next block horizontally, but also passed into the feature fusion module. This module was designed to enable the model to integrate both local and global information of gait. The module design is inspired by the network design of FPN (Feature Pyramid Network)[24]. FPN proposes a pyramid structure network to fuse feature map information of different scales. equently. These works Building upon FPN, various variant models[25–27] have been proposed subs highlight theutilization of feature maps of different scales, providing the model with greater flexibility in recognizing objects of varying sizes.

Inspired by the FPN network and related variant models, we introduce an feature fusion module to our model. This module performs feature map fusion in a top-down and bottom-up manner. To begin, the output of a convolutional block i, $f^i \in \mathrm{R}^{C\times H\times W}$, is processed by a 1 × 1 convolutional layer to obtain a rearranged result. Subsequently, the output is upsampled with a scale of 2 and added to the features from the previous block, completing the top-down fusion process.

$$f^i_{\text{top-down}} = \text{upsample}\left(\text{conv}\left(f^i\right)\right) \oplus \text{conv}\left(f^{i-1}\right) \quad (1)$$

Following that, in the bottom-up fusion process, a convolutional layer with a kernal size of 3 × 3 is applied to adjust the feature map. Next, the feature map from the bottom layer is processed and downsampled by a convolutional layer with a kernel size of 3 × 3 and a stride of 2. This down-sampled feature map is then added to the features of the subsequent block.

$$f^i_{\text{bottom-up}} = \text{downsample}\left(\text{conv}\left(f^i_{\text{top-down}}\right)\right) \oplus \text{conv}\left(f^i\right) \quad (2)$$

Finally, the fused feature map is completed using a 3 × 3 convolutional layer. By performing the above operations, the fused feature map of each convolutional block is obtained and fed into the next module.

The information fusion module enables the combination of local feature information at different scales, allowing the generation of global information from these local representations. As the fused feature maps of each convolutional block are interconnected with the previous and subsequent ones, diverse local information is reinforced, thus achieving a balance between local and global information weighting.

3.4 Dual attention module

The feature fusion module effectively integrates feature information from different blocks. To further enhance the fusion of local and global feature information, we introduce a dual attention module. This module consists of two consecutive attention mechanisms. The first one is dimension attention, which enhances feature representation from both semantic and spatial dimensions. The second one is scale attention, which selectively establishes the representation relationship between gait identity and various scale information.

3.4.1 Dimension attention

Dai et al.[28] proposed a multi-head detection method that combines self-attention mechanism in the object detection task, allowing the model to extract information by considering the attention from semantic, spatial, and task perspectives of the image. Similarly, in the gait recognition task, different semantics expressed by body parts and contours, as well as the spatial variations of different parts in the silhouette, are the information that the model needs to focus on and decouple. Therefore, to address this, the model in-corporates a dimension attention mechanism by integrating the design of self-attention.

After the preceding stages of information processing, the network generates set-level features denoted as $F \in \mathrm{R}^{L\times C\times H\times W}$, where L represents the stage dimension, and H and W represent the height and width of the feature map, respectively. For convenience in subsequent operations and representation, we can transform the two-dimensional representation into a one-dimensional graph representation. Let $I = H \times W$, and organize the set-level features as $F \in \mathrm{R}^{L\times C\times I}$. The overall process of the dimension attention mechanism can be represented by Eq. (3), where M_{sp} and M_{se} represent the attention information constructed in the spatial dimension and

the semantic dimension, respectively.

$$M(F) = M_{\mathrm{sp}}(M_{\mathrm{se}}(F) \cdot F) \cdot F \tag{3}$$

The semantic dimension represents different levels of semantic information in the images. By utilizing Eq. (4), it is possible to fuse feature information from different stages and extract semantics at different levels. In the equation, "Conv" represents a convolutional layer with a kernel size of 1 × 1, and $\sigma(x)$ denotes the hard-sigmoid function, which takes the value of $\max\left(0, \min\left(1, \frac{1}{2}\right)\right)$.

$$M_{\mathrm{se}}(F) \cdot F = \sigma\left(\mathrm{Conv}\left(\frac{1}{CI}\sum_{C,I} F\right)\right) \cdot F \tag{4}$$

To generate spatial information, a deformable convolutional layer[29] with a kernel size of 3 × 3 is applied to sparsely sample features at the same position of each graph. Here, n denotes the number of sampling regions, and these sampled features are aggregated in the semantic dimension. Δq_k represents the offset of the sampling region, and Δr_k is a trainable scalar.

$$M_{\mathrm{sp}}(F) \cdot F = \frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} \alpha_{l,k} \cdot S\left(l; q_k + \Delta q_k; c\right) \cdot \Delta r_k \tag{5}$$

The feature maps integrated through dimension attention mechanism enable the model to capture and learn information from different semantic and spatial dimensions.

3.4.2 Scale attention

Since human body information can be extracted in a block-by-block manner from top to bottom, pedestrian reidentification tasks usually perform slice processing on the feature maps. This method is known as Horizontal Pyramid Pooling (HPP), which was first proposed by Fu et al.[30]. Subsequent related works in gait recognition[6, 31] have also adopted this approach and made corresponding improvements. This work follows a similar approach to map the feature maps.

Specifically, for the feature map $F \in \mathrm{R}^{C\times H\times W}$ after integration through the dimension attention mechanism, we perform slice operations along the height dimension using the scale set $D = \{1, 2, 4, 8, ..., d\}$. After slicing, we obtain $\sum_1^d 2^{d-1}$ feature blocks. Let $f_{d,t}$ represent a feature block, where $t \in 1, 2, ..., 2d-1$ denotes the index under the d-scale slicing. Further compression processing is applied to these feature blocks, as shown in Eq. (6).

$$g_{d,t} = fc(\mathrm{mp}(f_{d,t}) + \mathrm{ap}(f_{d,t})) \tag{6}$$

The one-dimensional features $g_{d,t}$ are generated by processing the feature blocks through both the global max pooling layer (mp) and the global average pooling layer (ap), followed by integration using fully connected layers (fc).

The above operations compress the gait information in the network $\sum_1^d 2^{d-1}$ one-dimensional representations. These representations represent gait information at different partition scales, which can be considered as partitioning the image into different scales and concatenating the segmented image blocks. However, if the information is simply concatenated, the model cannot selectively learn the image block information that is more relevant to gait identity. Therefore, in this study, we introduce a multi-scale attention mechanism to analyze the interdependencies among different scale information, making the model's learning more efficient.

First, further compress the one-dimensional features, as shown in Eq. (7). After this processing, a feature descriptor is obtained for each information block. These descriptors are concatenated in order of scale to form a two-dimensional tensor $z \in \mathrm{R}^{d\times t}$ Then, the activation tensor e for each information block is generated using Eq. (8). FC1 and FC2 denote two fully connected layers with distinct parameters; ReLUis the rectified linear unit used to introduce non-linearity between different information blocks; σ is the sigmoid function, which ensures that the generated attention results are non-exclusive.

$$z_{d,t} = \mathrm{ap}(g_{d,t}) \tag{7}$$

$$e = \sigma(fc_2(\mathrm{ReLU}(fc_1(z)))) \tag{8}$$

Finally, the feature map with the scale-dependent relationships is obtained by element-wise multiplication between the activation results of information blocks and the original features, as shown in Eq. (9).

$$g'_{d,t} = e_{d,t} \otimes g_{d,t} \tag{9}$$

By learning this feature map representation, the model can obtain the correlations between different information blocks and establish a more selective relationship between gait identity and information at various scales.

3.5 Loss function

To improve the training efficiency of the model, a

hybrid loss function is used, which combines the triplet loss and cross-entropy loss.

The triplet loss function is commonly used in face recognition tasks, and its principle is to optimize model parameters based on the distances between positive and negative samples and an anchor sample. Assuming we have two gait sequences X^a and X^b from the same individual X, as well as another gait sequence Y^c from a different individual Y, the optimization objective of the triplet loss function is

$$L_{\text{tp}} = \left[\| f(X^a) - f\left(X^b\right) \|_2^2 - \| f(X^a) - f(Y^c) \|_2^2 + \alpha \right]_+ \quad (10)$$

In the equation, α is the margin parameter, $[]_+$ denotes the maximum taking function. When the function value is greater than zero, the output of the equation is equal to the function value; when the function value is equal to or less than zero, the output is 0. The objective of this function is to minimize the distance between samples of the same class while maximizing the distance between samples of different classes.

The cross-entropy loss function is commonly employed in classification tasks to quantify the dissimilarity between two distinct probability distributions of the same random variable. In the context of gait recognition task, we can interpret a person's gait behavior as the true probability distribution, and utilize the cross-entropy loss function to compute the disparity between the predicted probability distribution and the true probability distribution of the model's output. A smaller value indicates better prediction performance.

Specifically, for the application of the cross-entropy loss function, let N be the number of individuals in the training set. The predicted probability for the n-th individual is denoted as p_n, and its corresponding information entropy is denoted as h_n. The cross-entropy loss function can be expressed as follows:

$$L_{\text{ce}} = -\sum_{n=1}^{N} h_n \log p_n \quad (11)$$

Combining the above two loss functions, the final loss function used for training the model is given by

$$L = L_{\text{tp}} + L_{\text{ce}} \quad (12)$$

4 Experiment

4.1 Datasets

This model conducted relevant experiments on two major mainstream datasets: CASIA-B[32] and OU-MVLP[33].

CASIA-B dataset. The CASIA-B dataset comprises gait experimental data from 124 subjects, recorded from various angles. The data includes RGB images and silhouette images,

making it one of the most widely used datasets in gait recognition research.

To capture data from multiple angles, the dataset was recorded at 11 different angles with an interval of 18°, ranging from 0° to 180°. The dataset is categorized into three walking conditions: normal walking (NM), walking with a coat (CL), and walking with a bag (BG). For each of these conditions, each subject at each angle has 6, 2, and 2 samples, respectively.

In accordance with standard experimental procedures, the initial 74 subjects' data is employed for training, while the remaining 50 subjects' data is reserved for testing. During the testing phase, each subject and angle's first 4 samples of normal walking serve as the reference set, while the subsequent 6 samples are utilized as the validation set.

OU-MVLP dataset. The OU-MVLP dataset is currently the largest in terms of the number of samples, containing a total of 259 013 gait samples from 10 307 subjects. The age distribution of the subjects ranges from 2 to 87 years, evenly distributed. The gait data was recorded at angles from 0° to 90° and from 180° to 270°, with an interval of 14°, and each angle has two sets of data.

As per the data provider's guidelines, the data from the first 5153 subjects is designated as the training set, while the data from the remaining 5154 subjects is used as the test set. For the testing phase, the reference set is comprised of the initial set of data captured from each subject at every angle, and the validation set comprises the second set of data.

4.2 Experimental settings

In this study, all experimental data inputs consist of gait silhouette images with a resolution of 64 × 44. The length of each gait input sequence is fixed at 30 frames. When training on the CASIA-B dataset and the OU-MVLP dataset, the number of output channels in the model's convolutional layers is configured as follows: 64, 128, 256, and 512 for the four convolutional blocks, respectively. The feature map slicing operation employs a scale set denoted as D, defined as 1, 2, 4, 8, and 16.

During model training, the initial learning rate is set to 10^{-2}, and a momentum update of 0.9 is applied. Additionally, a weight decay rate of 5×10^{-4} is utilized.

4.3 Ablation study

To further evaluate the efficacy of the designed modules in our model, we conducted ablation experiments using the CASIA-B dataset.

4.3.1 Dimension attention

The introduction of the dimension attention mechanism enhances the model's ability to capture different semantic and spatial information. In the baseline model, we added the corresponding attention module to perform ablation comparison. By incorporating the dimension attention into the baseline model, we evaluated how it contributed to the model's performance in capturing and emphasizing important features related to gait recognition.

To validate the effectiveness of the module design and the rationale behind the parameter settings, we performed relevant ablation experiments on the CASIA-B dataset.

As shown in Table 1, when introducing only the semantic attention, baseline model achieved an accuracy improvement of 0.3%, 0.7%, and 1.7% on NM, BG, and CL benchmark tests, respectively. On the other hand, when introducing only the spatial attention, baseline model achieved an accuracy improvement of 0.5%, 1.1%, and 1.4% on NM, BG, and CL benchmark tests, respectively. It is worth noting that, in the BG benchmark, the improvement from spatial attention was higher than that from semantic attention. Considering the data characteristics and module design analysis, the bag only affects the local attributes of the subjects, while the spatial attention enables the model to focus on more relevant parts of the gait, thus avoiding interference caused by the bag.

In contrast, in the CL benchmark, the coat wearing setting changes the global appearance attributes of the subjects, and semantic attention can handle semantic information at the level of the gait silhouette more effectively. Therefore, it provides a higher gain for the model. When combining both types of attention and introducing them into the Baseline model, the accuracy improved by 0.9%, 1.4%, and 1.9% in the NM, BG, and CL benchmark tests, respectively. The experimental results indicate that the dimensional attention mechanism can help the model better disentangle gait information and improve overall robustness.

Table 1 Experiment on dimension attention.

Model	NM (%)	BG (%)	CL (%)
Baseline	95.2	88.0	71.0
Baseline + semantic	95.5	88.7	72.7
Baseline + spatial	95.7	89.1	72.4
Baseline + both	96.1	89.4	72.9

4.3.2 Scale attention

The experiment on scale attention mainly focuses on two aspects: the selection of feature map slicing and the effectiveness of the attention module in improving model accuracy.

Firstly, we explore the feature map slicing. The scale set $D = \{1, 2, 4, 8, ..., d\}$ determines the number and size of the sliced feature blocks. Therefore, to find the appropriate set, the feature map is sliced according to different scale sets, and experimental results are tested. The scale set settings and experimental results are shown in Table 2.

Due to the relationship: $\sum_1^d 2^{d-1}$ between the number of sliced feature blocks and the value of d in the scale set, a larger value of d will result in a higher number of feature representations with smaller dimensions From the experimental results, we observed that using smaller slicing scales leads to lower discriminative power between features, as the information becomes more sparse, making it difficult for the model to learn nonlinear relationships. On the other hand, larger slicing scales result in a larger feature space, making it challenging to form compact representations and leading to a decrease in model performance.

Investigating the performance of the attention module is another aspect of the study. After slicing the feature map with the scale set {1, 2, 4, 8, 16}, the sliced results are passed through a fully connected layer for integration, and then fed into the attention module for further processing.

As depicted in Table 3, the incorporation of the multiscale attention module led to notable accuracy

Table 2 Experiment on feature map slicing scale set.

Set name	Slicing scale	NM (%)	BG (%)	CL (%)
D1	1	95.2	88.0	71.0
D2	1, 2	95.5	88.7	72.7
D3	1, 2, 4	95.7	89.1	72.4
D4	1, 2, 4, 8	96.1	89.4	72.9
D5	1, 2, 4, 8, 16	95.7	89.1	72.4
D6	1, 2, 4, 8, 16, 32	96.1	89.4	72.9

Table 3 Experiment on scale attention.

Model	NM (%)	BG (%)	CL (%)
Baseline	95.2	88.0	71.0
Baseline+scale Att.	95.7	90.0	73.1

improvements of 0.5%, 2.0%, and 2.1% on the NM, BG, and CL benchmarks, respectively, for the Baseline model. This module can capture the inter-scale dependencies and emphasize the expression of information that is more relevant to gait, making the learning and modeling process of the model more efficient.

4.4 Model performance comparison

CASIA-B. Table 4 presents a comparison of the accuracy achieved by various models on the CASIA-B dataset across three different testing benchmarks.

We have included a selection of representative methods for conducting our experimental comparisons. Among these, GEINet serves as a representative of template-based approaches, whereas CNN-3D, CNN-Ensemble, CNN-LB, and Gaitset represent sequence-based methods. These methods employ distinct modeling approaches, each characterized by unique network structures. The first three methods all operate on ordered sets as their input. CNN-3D employs a 3D CNN architecture to enhance the analysis of temporal information. CNN-Ensemble, on the other hand, leverages the integration of multiple independent CNN models to enhance predictive performance. CNN-LB combines features from both CNN and LSTM architectures. Conversely, Gaitset introduces a novel concept of processing unordered sets as input, utilizing a multi-layer CNN network to dissect and fuse set features. This approach has since served as a source of inspiration for numerous subsequent gait-related methods.

Based on the data presented in the table, it is evident that in multi-angle recognition, the angles 0° and 180° exhibit the lowest recognition accuracy, closely followed by 90°. Analyzing the data characteristics, 0° and 180° correspond to frontal and back views of individuals, respectively. The swinging amplitude and frequency of limbs are difficult to observe from these angles, resulting in limited gait information compared to other angles. As for 90°, it represents the lateral view, and one side of the limbs will obstruct the other side, making it challenging to obtain a complete gait information from this angle. The more ideal gait observation angles are oblique lateral views, such as 36°and 126°.

Under all three testing benchmarks, our proposed method achieves an average accuracy surpassing other models, and it also outperforms them in almost all testing angles. This demonstrates that our model has excellent recognition performance and a certain level of robustness.

OU-MVLP. Based on the experimental findings presented in Table 5, angles 0° and 180° also pose significant challenges for recognition in this dataset. Our model outperforms other models in almost all testing angles. Moreover, OU-MVLP covers a wider

Table 4 Performance comparison based on CASIA-B dataset.

(%)

Benchmark	Model	Angle											Average
		0°	18°	36°	54°	72°	90°	108°	126°	134°	162°	180°	
NM	GEINet[34]	40.2	38.9	42.9	45.6	51.2	42.0	53.5	57.6	57.8	51.8	47.7	48.1
	CNN-3D[4]	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
	CNN-Ensemble[4]	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[6]	90.1	97.8	99.0	96.7	93.1	91.1	94.1	97.5	97.9	96.2	**87.0**	95.1
	Ours	**92.4**	**98.4**	**99.2**	**97.4**	**95.1**	**93.2**	**95.9**	**98.1**	98.0	**97.3**	86.3	**95.6**
BG	GEINet	34.2	29.3	31.2	35.2	35.2	27.6	35.9	43.5	45.0	38.9	36.8	35.7
	CNN-LB[4]	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	84.1	92.5	92.7	91.0	85.6	80.1	84.1	91.2	92.9	91.1	80.3	88.8
	Ours	**87.6**	**95.0**	**95.1**	**93.7**	**89.6**	**85.3**	**90.3**	**93.5**	**95.3**	**92.4**	**82.5**	**90.9**
CL	GEINet	19.9	20.3	22.5	23.5	26.7	21.3	27.4	28.2	24.2	22.5	21.6	23.5
	CNN-LB	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	63.1	74.4	81.1	74.1	69.0	67.4	68.2	74.4	73.8	69.1	56.5	71.6
	Ours	**66.0**	**80.3**	**82.4**	**77.8**	**73.6**	**71.1**	**73.5**	**78.90**	**77.90**	**70.2**	**59.2**	**73.7**

Table 5 Comparison of overall performance of models based on the OU-MVLP dataset.

(%)

Model	Angle														Average
	0	15	30	45	60	75	90	180	195	210	225	240	255	270	
GEINet	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	44.9	45.9	45.7	41.0	42.5
GaitSet	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	**86.5**	89.0	89.2	87.2	87.6	86.2	87.1
Ours	**83.3**	**88.6**	**90.7**	**90.5**	**89.5**	**89.1**	**88.4**	**84.6**	86.1	**89.7**	**89.3**	**88.9**	**88.2**	**86.8**	**88.1**

range of recognition angles, which demands higher adaptability from the model to handle angle variations, making its results more convincing.

5 Conclusion

This research introduces a gait recognition model based on feature fusion and dual attention to address challenges in the field. The model utilizea a ResNet-based network for feature extraction. A feature fusion module analyzes gait local features to create an effective global feature representation. To capture different semantics and spatial information, the model employs dual attention mechanisms. The dimension attention mechanism captures information from different semantic layers and spatial locations. The scale attention mechanism analyzes interdependencies between segmented information at different scales.

The model is trained using a hybrid approach, incorporating both triplet loss and cross-entropy loss functions to learn discriminative features. Experiments conducted on CASIA-B and OU-MVLP datasets showcase the effectiveness of the dual attention mechanism, and the proposed model surpasses other methods in performance across vari- ous benchmarks.

Our future work will be dedicated to improving accuracy and addressing model deficiencies, as well as exploring unsupervised methods for gait recognition to handle unlabeled data.

Acknowledgment

This work was supported by the Fujian Provincial Department of Science and Technology, and 2022 Fuxiaquan Autonomous Innovation Demonstration Zone Collaborative Innovation Platform Project (No. 2022FX6).

References

[1] Y. W. He, J. P. Zhang, H. M. Shan, and L. Wang, Multi-task gans for view-specific feature learning in gait recognition, *IEEE Transactions on Information Forensics and Security*, vol. 14, pp. 102–113, 2019.

[2] M. Hu, Y. Wang, Z. Zhang, J. J. Little, and D. Huang, View-invariant discriminative projection for multi-view gait-based human identification, *IEEE Trans. Inf. Forensics Secur.*, vol. 8, no. 12, pp. 2034–2045, 2013.

[3] 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, 2019.

[4] 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, 2017.

[5] J. G. Han and B. Bhanu, Individual recognition using gait energy image, *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, 2005.

[6] H. Q. Chao, Y. W. He, J. P. Zhang, and J. F. Feng, Gaitset: Regarding gait as a set for cross-view gait recognition, arXiv preprint arXiv:1811.06186, 2019.

[7] Y. Zhang, Y. Huang, S. Yu and L. Wang, Cross-View Gait Recognition by Discriminative Feature Learning, *IEEE Transactions on Image Processing*, vol. 29, pp. 1001–1015, 2020.

[8] C. Fan, Y. J. Peng, C. S. Cao, X. Liu, S. H. Hou, J. N. Chi, Y. Z. Huang, Q. Li, and Z. Q. He, Gaitpart: Temporal part-based model for gait recognition, in *Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, pp. 14213–14221, 2020.

[9] X. H. Wu, W. Z. An, S. Q. Yu, W. Y. Guo, and E. B. G. Reyes, Spatial-temporal graph attention network for video-based gait recognition, in *Proc. ACPR*, Auckland, New Zealand, 2019.

[10] R. J. Liao, C. S. Cao, E. B. G. Reyes, S. Q. Yu, and Y. Z. Huang, Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations, in *Proc. CCBR*, Shenzhen, China, 2017.

[11] G. Ariyanto, Model-based 3d gait biometrics, in *Proc. 2011 Int. Joint Conf. on Biometrics* (*IJCB*), Washington, DC, USA, pp. 1–7, 2011.

[12] Y. Liu, X. Jiang, T. F. Sun, and K. Xu, 3d gait recognition based on a cnn-lstm network with the fusion of skegei and da features, in *Proc. 2019 16th IEEE Int. Conf. on Advanced Video and Signal Based Surveillance* (*AVSS*), Taipei, China, pp. 1–8, 2019.

[13] T. Huynh-The, C.-H. Hua, N. A. Tu, and D.-S. Kim, Learning 3d spatiotemporal gait feature by convolutional network for person identification, *Neuro-Computing*, vol. 397, pp. 192–202, 2020.

[14] F. Ahmed, P. P. Paul, and M. Gavrilova, Kinect-based gait

recognition using sequences of the most relevant joint relative angles, https://dspace5.zcu.cz/bitstream/11025/17149/1/Ahmed.pdf, 2015.

[15] Y. Feng, Y. C. Li, and J. B. Luo, Learning effective gait features using lstm, in *Proc. 2016 23rd Int. Conf. on Pattern Recognition* (*ICPR*), Cancun, Mexico, pp. 325–330, 2016.

[16] C. Wang, J. P. Zhang, J. Pu, X. R. Yuan, and L. Wang, Chrono-gait image: A novel temporal template for gait recognition, in *Proc. European Conf. on Computer Vision*, Crete, Greece, 2010.

[17] M. H. Ghaeminia and S. B. Shokouhi, Gsi: Efficient spatio-temporal template for human gait recognition, *Int. J. Biom.*, vol. 10, pp. 29–51, 2018.

[18] K. Bashir, T. Xiang, and S. G. Gong, Feature selection on gait energy image for human identification, in *Proc. 2008 IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Las Vegas, NV, USA, pp. 985–988, 2008.

[19] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and A. Bouridane, Gait recognition for person re-identification, *J. Supercomput.*, vol. 77, no. 4, pp. 3653–3672, 2021.

[20] B. Lin, S. L. Zhang, and F. Bao, Gait recognition with multiple-temporal-scale 3d convolutional neural network, in *Proc. of the 28th ACM Int. Conf. Multimedia*, New York, NY, USA, 2020.

[21] G. H. Huang, Z. Lu, C. M. Pun, and L. L. Cheng, Flexible gait recognition based on flow regulation of local features between key frames, *IEEE Access*, vol. 8, pp. 75381–75392, 2020.

[22] R. J. Liao, S. Q. Yu, W. Z. An, and Y. Z. Huang, A model-based gait recognition method with body pose and human prior knowledge, *Pattern Recognit.*, vol. 98, p. 107069, 2020.

[23] C. Fan, J. H. Liang, C. F. Shen, S. H. Hou, Y. Z. Huang, and S. Q. Yu, Opengait: Revisiting gait recognition towards better practicality, in *Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Vancouver, Canada, pp. 9707–9716, 2023.

[24] T.-Y. Lin, P. Dollár, R. B. Girshick, K. M. He, B. Hariharan, and S. J. Belongie, Feature pyramid networks for object detection, in *Proc. 2017 IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Honolulu, HI, USA, pp. 936–944, 2017.

[25] S. Liu, L. Qi, H. F. Qin, J. P. Shi, and J. Y. Jia, Path aggregation network for instance segmentation, in *Proc. 2018 IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 8759–8768, 2018.

[26] M. X. Tan, R. M. Pang, and . V. Le, Efficient-det: Scalable and efficient object detection, in *Proc. 2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Seattle, WA, USA, pp. 10778–10787, 2020.

[27] S. Y. Qiao, L.-C. Chen, and A. L. Yuille, Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution, in *Proc. 2021 IEEE Conf. on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, TN, USA, 2021.

[28] X. Y. Dai, Y. P. Chen, B. Xiao, D. D. Chen, M. C. Liu, L. Yuan, and L. Zhang, Dynamic head: Unifying object detection heads with attentions, in *Proc. IEEE/CVF conference on computer vision and pattern recognition*, Nashville, TN, USA, pp. 7373–7382, 2021.

[29] J. F. Dai, H. Z. Qi, Y. W. Xiong, Y. Li, G. D. Zhang, H. Hu, and Y. C. Wei, Deformable convolutional networks, in *Proc. 2017 IEEE Int. Conf. on Computer Vision* (*ICCV*), Venice, Italy, pp. 764–773, 2017.

[30] Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. Huang, Horizontal pyramid matching for person re-identification, *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, pp. 8295–8302, 2019.

[31] S. H. Hou, C. S. Cao, X. Liu, and Y. Z. Huang, Gait lateral network: Learning discriminative and compact representations for gait recognition, in *Proc. European Conf. on Computer Vision*, Glasgow, UK, pp. 382–398, 2020.

[32] 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 Recognition-Volume 04*, Cambridge, UK, 2006, pp. 441–444.

[33] 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 Transactions on Computer Vision and Applications*, vol. 10, pp. 1–14, 2018.

[34] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, Geinet: View-invariant gait recognition using a convolutional neural network, in *Proc. 2016 Int. Conf. on Biometrics* (*ICB*), Halmstad, Sweden, pp. 1–8, 2016.

Zhixiong Wu is a PhD student in Department of Computer Science and Technology at Tsinghua University, China. He is the CEO of Linewell Software Group, vice chairman of the China Software Industry Association, president of the Fujian Software Industry Association, and deputy director of the Management Committee of Tsinghua University and Linewell Software Digital Governance Information Technology Joint Research Center, a national high-level talent in science and technology. His major research interests include artificial intelligence, data elements, blockchain, and big data intelligence.

Yong Cui received the BE and PhD degrees both in Computer Science and Engineering from Tsinghua University, China. He is currently a full professor with Computer Science Department in Tsinghua University. He served or serves at the editorial boards on *IEEE TPDS*, *IEEE TCC*, *IEEE Network*, and *IEEE Internet Computing*. He published over 100 papers with several Best Paper Awards and 10 Internet standard documents (RFC). His research interests include Internet architecture and data-driven network.