Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18692-0

AdaptiveGait: adaptive feature fusion network for gait recognition

Tian Liang[1] · Zhenxue Chen[1] · Chengyun Liu[1] · Jiyang Chen[2,3] · Yuchen Hu[1] · Q. M. Jonathan Wu[4]

Received: 10 December 2023 / Revised: 21 February 2024 / Accepted: 23 February 2024

Abstract
Gait recognition is a biometric approach used to identify people based on their walking patterns at long distances and low resolutions. Most advanced gait recognition methods based on silhouettes employ the focal convolution module. However, experiments have demonstrated that the horizontal segmentation method used in this module causes information loss at the feature map demarcation line. In this paper, we propose an adaptive feature fusion block (AFFB) for feature extraction that utilizes comprehensive global features to compensate for the lost local features, significantly reducing feature loss caused by local convolution. Additionally, we introduce a feature expansion module (FEM) to enrich the temporal information of gait features and adaptively balance the body detail information extracted by the model with the overall body information . We evaluated our model on CASIA-B and OU-MVLP datasets and compared it to other gait models using RANK-1 accuracy. The experimental results show that our model can represent gait features better than other models and achieved high accuracy in gait recognition across perspectives and various walking conditions.The source code will be available on https://github.com/Lentia/AdaptiveGait.

Keywords CNN · Gait recognition · Adaptive feature fusion · Feature expansion · Cross-view

1 Introduction

Gait recognition is an individualized recognition method that identifies a person's distinctive walking pattern. It distinguishes itself from other biometric techniques, such as face, iris, or fingerprint recognition by allowing for contactless, long-range, and low-resolution recogni-

Tian Liang and Zhenxue Chen contributed equally to this work

✉ Zhenxue Chen
chenzhenxue@sdu.edu.cn

✉ Chengyun Liu
liuchengyun@sdu.edu.cn

Extended author information available on the last page of the article

Published online: 11 March 2024

tion. Given that gait recognition does not require the active cooperation of the subject, it has broad potential in crime prevention, forensic identification, and social security.

Various methods have been developed since gait recognition was introduced, and many have achieved good performance. However, there some issues with existing methods still exist, considering that recognition accuracy is considerably impacted by a person's clothes, carrying situation, viewpoint, and other factors. For example, Fig. 1 illustrates the gait silhouettes of a pedestrian wearing a backpack or coat. These factors can have a substantial effect on the accuracy of gait recognition.

Many deep-learning-based gait recognition methods have been generated in recent years due to better performance in terms of accuracy and sophistication compared to traditional methods, and some of these methods extract global or local features from gait silhouettes. For example, Chao et al. [2] used 2D CNN to extract global gait features in gait sequences. Fan et al. [3] proposed focal convolution to extract features for different parts of the human body. Lin et al. [13] proposed a Global and Local Feature Extractor (GLFE) to extract both global and local information from gait silhouettes. Global features predominantly contain more spatial and temporal gait information, and local features focus more on the spatial and temporal information of different parts of the body. Both global and local information contributes significantly to the effectiveness of gait recognition. Hence, the adequate extraction of both gait features is a crucial aspect of gait recognition.

However, most existing methods rely on a horizontal splitting of the feature map in the process of local information extraction. This leads to the gait features of various parts of the body being concentrated into horizontally divided regions and "chunking" of features. The feature maps at the boundaries of the chunked regions and the connections between different parts of the body often appear significantly weakened or may even disappear, which can greatly affect the gait feature representation.

This paper proposes an adaptive feature fusion block (AFFB) that can incorporate global features and compensate for the absent parts in local features. Through such a method, the extracted information can be supplemented adaptively, leading to a more comprehensive gait representation. Furthermore, given that feature map chunking results in weaker connections between body parts, it may hinder the extraction of the temporal information of the gait as a whole. To avoid focusing on local gait information in the feature extraction component, we will design and utilize a feature expansion module (FEM) at the end of the feature extraction process. FEM uses a conventional 3D convolutional block to retain more spatial and temporal gait information and amplifies the gait features in the channel dimension to obtain a more comprehensive representation. Finally, the expanded features are mapped to the feature space by adaptive horizontal pooling, with the addition of a fully connected layer of the final gait features.

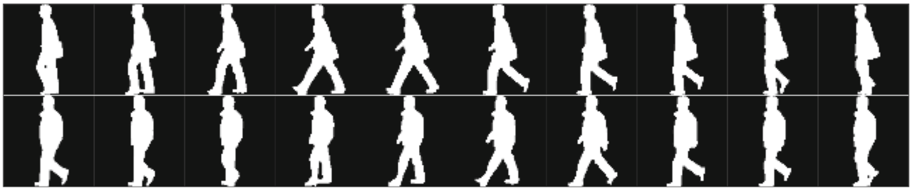


Fig. 1 Gait silhouettes from the CASIA-B [31] dataset. The first row shows the gait silhouette in the backpack state, and the second row shows the gait silhouette in the coat-wearing state. It can be observed that the gait contours of pedestrians in the backpack and coat-wearing states appear to change significantly, which affects the extraction of gait features

The primary components of this work can be summarized as follows:

- We proposed a novel adaptive feature fusion method for gait feature extraction, which can effectively balance the ratio of global and local features and render the extracted gait features more comprehensively.
- We designed a feature expansion module which enhances the temporal and spatial information of gait features, expands the extracted body detail information and overall information, and enhances the expressiveness of the features.
- Our network was tested on two commonly used gait datasets (CASIA-B [31], OU-MVLP [21]), and the experimental results demonstrate the effectiveness of our method and its competitiveness with existing advanced gait recognition methods.

The rest of this paper is structured as follows: Firstly, related works are summarized in Section 2. In Section 3, we describe the key modules of our network and in Section 4, we present the details of the comparative and ablation experiments. In Section 5, we discuss the experimental results and future work.

2 Related work

2.1 Gait recognition

Two gait recognition methods have been proposed to address the challenges posed by cross-view and various walking conditions, namely the model-based approach and the appearance-based approach. Template-based methods [10, 12, 23] usually model the basic structure of human posture through the use of gait images and use the parameters of the human model as features for recognition. For example, Liao et al. [12] proposed PoseGait, which uses convolutional neural networks to estimate the 3D poses of pedestrians from viable gait images, and extracts effective features of the gait from the 3D poses of pedestrians. Li et al. [10] proposed a model-based, end-to-end gait recognition method that uses a multi-person linear (SMPL) model with skin for human modeling. Teepe et al. [23] proposed GaitGraph for obtaining accurate human poses, combining skeleton poses with a graph convolutional network (GCN) to estimate human skeleton poses from RGB images. Template-based methods have better robustness due to their direct modeling of human body structure, and they are less affected by covariates such as inter-view variations. However, model-based methods are computationally intensive and the accuracy of the generated human models significantly impacts the accuracy of the final recognition.

Many appearance-based methods have been proposed since these are less computationally intensive and easier to train and use than template-based methods. Shiraga et al. [20] aggregated gait silhouettes into a gait energy image (GEI) and used a 2D convolutional network to extract gait features from it, Xu et al. [26] also used the GEI method for gait recognition and proposed a pairwise spatial transformer network (PSTN) to reduce the feature mis-alignment caused by view differences, thus improving the recognition performance; however, a large amount of fine-grained spatial information was lost in the GEI aggregation process. To address this issue, Chao et al. [2] proposed GaitSet, which treats the gait silhouette as an unordered set and extracts gait features from the gait silhouette using a 2D CNN. Fan et al. [3] proposed GaitPart, which divides the feature map horizontally to extract spatial and temporal features separately from different parts of the body in the gait silhouette using the focal convolution module (FConv). Hou et al. [5] proposed the Gait Lateral Network (GLN) to enhance gait representation using the inherent feature pyramid in deep convolutional neural networks. Qin et al. [18] noticed the problem of horizontal partitioning of feature maps and strengthened the connections between each block by analyzing the relationships between different parts of the gait feature map. However, they did not extract the temporal information well. Lin et al. [13] proposed Global and Local Feature Extractor (GLFE) by exploiting the superiority of 3D-CNN in dealing with sequential problems and combining the idea of FConv, achieving significant improvements in recognition performance. However, this method ignored the proportional relationship between global and local information in the feature extraction process. Huang et al. [6] proposed the Spatial-Temporal Dual-Attention (STDA) unit by combining the idea of 3D convolutions with spatial-temporal decoupling, in order to better utilize the temporal and spatial information of gait. Chai et al. [1] proposed LagrangeGait, which combines motion extraction and viewpoint embedding, and achieves excellent detection results.

From the analysis of appearance-based methods, it can be seen that extracting complete gait features is a vital part of gait recognition. In this paper, we propose an adaptive feature fusion block (AFFB) to adaptively compensate for the loss of local features and improve the gait features extracted. Furthermore, we design a feature expansion module (FEM) to further enhance the temporal and spatial features of gait, thus obtaining a more comprehensive gait feature representation.

2.2 Adaptive feature fusion

Adaptive feature fusion is an effective method for feature combination. In networks, using only a single feature is insufficient to express the complexity of a problem, since the output features of different layers contain specific feature information. For example, the shallow features of convolutional neural networks tend to have abundant geometric details, while the deeper features concentrate more on abstract semantic information. As such, adaptive feature fusion has been extensively employed in computer vision tasks, such as semantic segmentation [14, 19], target detection [7, 8], vehicle detection [24], multi-view learning [27–29] and other applications [11, 16, 17, 30]. Numerous methods exist for feature fusion, including feature summation and stitching, and element-wise multiplication. Based on these methods, multiple adaptive feature fusion strategies have been proposed.

Liu et al. [15] proposed Path Aggregation Network (PANet), which utilizes a feature pyramid network to fuse different feature hierarchies effectively. Subsequently, Zhao et al. [32] proposed m2det, which uses a multi-level feature pyramid network to further improve detection accuracy. Ghiasi et al. [4] proposed a novel feature pyramid network structure (NAS-FPN), which enables feature fusion at multiple scales. Moreover, Tan et al. [22] proposed a Weighted Bi-directional Feature Pyramid Network (BiFPN) to integrate features of different scales using trainable weights. Xu et al. [28] achieved dynamic and adaptive multimodal fusion at the evidence level, synthesizing information from multiple views to make reliable predictions. [27] and [29] explored multi-modal learning methods to guarantee the consistency and complementarity properties in the multi- frequency data, providing valuable insights for related work on adaptive feature fusion.

Unlike the above methods, we do not propose fusing the features of different layers in this paper, as the gait silhouette is a set of binary images containing limited information. Fusing the shallow features with the deep features may introduce some negative information, thus making gait recognition less effective.

For the same-layer features of gait recognition, the global features capture overall, coarse-grained, and temporal information, while the local features focus on local, fine-grained, and spatial information. Therefore, in the process of extracting gait features, attention should be paid to the proportion of global features and local features, so that gait features can be expressed comprehensively and sufficiently.

3 Methods

In this section, we will first outline the construction of the proposed method. Then, we will describe the model's key components, including the AFFB and the FEM. Finally, the details of training and testing will be presented.

3.1 Overview

Our model is depicted in Fig. 2. It consists of four modules, i.e., Local Temporal Aggregator (LTA), Adaptive Feature Extractor, Feature Expansion Module (FEM), and Horizontal Pyramid Pooling (HPP). First, a gait sequence is inputted into the model, and the shallow features of the sequence are extracted using 3D convolution. The extracted shallow features are compressed in the temporal dimension using Local Temporal Aggregation to maximize the retention of temporal information. Next, the global and local features are synthesized with AFFB to extract more comprehensive gait features. The gait features obtained by the adaptive feature extractor are then fed into the FEM to balance the body detail information captured by the model with the body information. Finally, the gait features are mapped using HPP, and the model is trained with Triplet Loss and Cross-Entropy Loss. Among these, AFFB and FEM are the key techniques of this study, and will be further discussed in the following sections. Meanwhile, the same methods as [13] are adopted for LTA and HPP.

3.2 Adaptive feature fusion block (AFFB)

We designed an adaptive feature extractor based on the adaptive feature fusion block (AFFB). The feature extractor comprises two CNN blocks and a maximum pooling layer with a specific structure of AFFB-Max Pooling-AFFB, which can extract temporal and spatial information from gait sequences after temporal aggregation. In related research on gait recognition, GaitSet [2] is a typical model for extracting global gait features using 2D-CNN. In contrast,

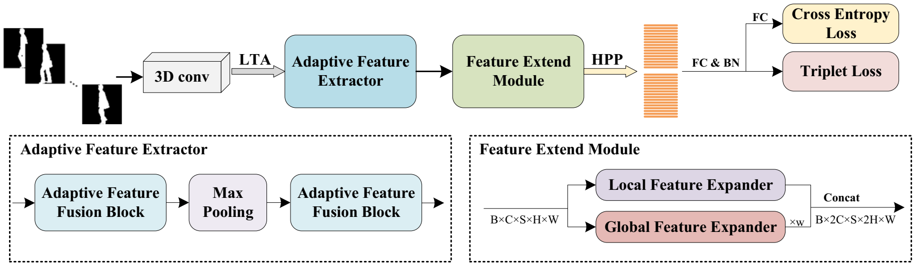


Fig. 2 Framework of the adaptive feature fusion gait recognition network. $B \times C \times S \times H \times W$ represents the dimension of the feature map. B is the batch size, C is the number of channels, S is the length of the feature map sequence, and (H, W) is the size of the feature map. W in the feature extend module is a trainable parameter

GaitPart [3] uses its proposed FConv to extract local information, showing that local features are essential in the gait recognition process. GaitGL [13] combines the ideas of [2] and [3], considering the use of global and local information, and using 3D-CNN for better temporal information extraction from gait sequences. Although GaitGL is aware of the problem of "chunking" of the feature map due to horizontal segmentation, simply adding global features to local features is still insufficient to compensate for the problems caused by chunking. In this paper, we propose AFFB to extract gait features that can make the global and local features of gait balance adaptive, thus compensating for the missing parts in the chunked features. The concrete implementation of AFFB is shown in Fig. 3.

First, AFFB contains two 3D convolution blocks: a global feature extraction block, and a local feature extraction block. As shown in Fig. 3, we use a 3D convolution to extract global information from the gait sequence in the global feature extraction block. In the local feature extraction block, a 3D convolution is applied to different parts of the feature map using the idea of horizontal segmentation. The 3D convolutions in the local feature extraction block share the same weights. We multiply the global features using a trainable weight "w_{AFFB}" and then add them to the local features using the residual structure. This method allows the model to autonomously find the best fusion weight of global and local features during the training process to achieve the complement of missing features. The final formed features will also be more detailed in terms of temporal and spatial information.

The gait features for each frame are calculated as follows: assume that the input of AFFB is $X_{whole} \in \mathbb{R}^{C_{in} \times S \times H \times W}$, where C_{in} is the number of input channels of AFFB, S is the length of the feature map sequence, and (H, W) is the size of each frame of the feature map. In order to facilitate the expression of the local feature extraction module, we horizontally divide each frame of the input feature map level into n parts, remember $X_{part} = \{X^i_{part} \mid i = 1, 2, ...n\} \in \mathbb{R}^{C_{in} \times S \times \frac{H}{n} \times W}$. The 3D convolution of global feature extractor is denoted as $f^{3\times3\times3}_{global}(\cdot)$, and the 3D convolution of local feature extractor is denoted as $f^{3\times3\times3}_{local}(\cdot)$. Then,

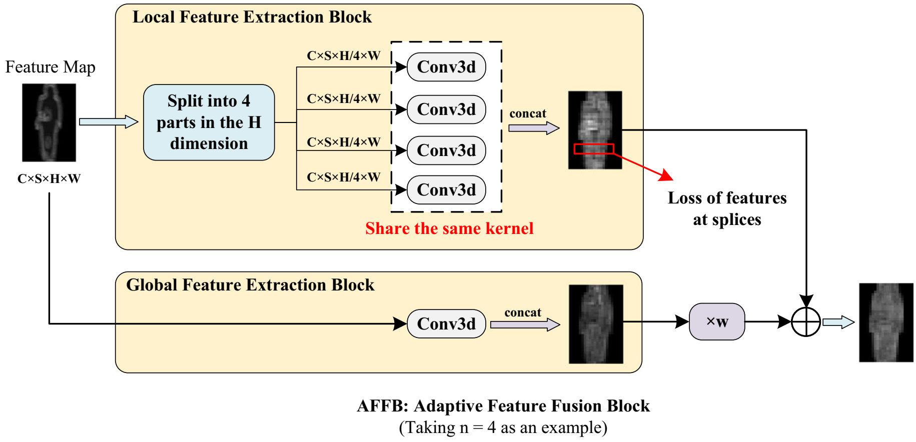

Local Feature Extraction Block
Feature Map
C×S×H×W
Split into 4 parts in the H dimension
C×S×H/4×W
C×S×H/4×W
C×S×H/4×W
C×S×H/4×W
Conv3d
Conv3d
Conv3d
Conv3d
concat
Share the same kernel
Loss of features at splices
Global Feature Extraction Block
Conv3d
concat
×w
AFFB: Adaptive Feature Fusion Block
(Taking n = 4 as an example)

Fig. 3 The illustration of AFFB, with C × S × H × W as the dimensions of the feature maps. The local feature extraction block uses the idea of focal convolution [3]. Here is an example of dividing the feature map into four blocks horizontally, with each convolution kernel sharing the same parameters. We use a 3D convolution as a global feature extraction block. W is a trainable parameter

the global feature and local feature extracted by AFFB can be expressed as:

$$Y_{global} = f_{global}^{3\times3\times3}(X_{whole}) \in \mathbb{R}^{C_{out}\times S\times H\times W}, \tag{1}$$

$$Y_{local} = cat\left\{\begin{array}{c} f_{local}^{3\times3\times3}(X_{part}^{1}) \\ f_{local}^{3\times3\times3}(X_{part}^{2}) \\ \cdots \\ f_{local}^{3\times3\times3}(X_{part}^{n}) \end{array}\right\} \in \mathbb{R}^{C_{out}\times S\times H\times W}, \tag{2}$$

where C_{out} is the number of output channels of AFFB.

Based on the above expressions for global and local features, the output of AFFB can be expressed as:

$$Y_{AFFB} = w_{AFFB}\cdot Y_{global} + Y_{local} \in \mathbb{R}^{C_{out}\times S\times H\times W}, \tag{3}$$

where w_{AFFB} is a trainable parameter.

The specific parameters of AFFB in the model are shown in Table 1.

3.3 Feature expansion module (FEM)

The feature extending module (FEM) aims to extend the original gait features into more scales, allowing more comprehensive gait information to be captured in the feature space. Its specific architecture is shown in Fig. 4. The gait features obtained by ordinary three-dimensional convolution include more temporal and spatial gait information, which is lacking in the focal convolution layer. After extracting the gait features, FEM splices the output of the

Table 1 The specific configurations of the proposed method under two different datasets

Layer	In_C	Out_C	Kernal	w	N-part
CASIA-B					
Conv3d	1	32	(3, 3, 3)	-	-
LTA	32	32	(3, 1, 1)	-	-
AFFB	32	64	(3, 3, 3)	11	4
Max Pooling	-	-	(1, 2, 2)	-	-
AFFB	64	128	(3, 3, 3)	11	8
FEM	128	256	(3, 3, 3)	6	16
OU-MVLP					
Conv3d	1	32	(3, 3, 3)	-	-
Conv3d	32	32	(3, 3, 3)	-	-
LTA	32	32	(3, 1, 1)	-	-
AFFB	32	64	(3, 3, 3)	3	4
AFFB	64	64	(3, 3, 3)	3	4
Max Pooling	-	-	(1, 2, 2)	-	-
AFFB	64	128	(3, 3, 3)	3	4
AFFB	128	128	(3, 3, 3)	3	4
AFFB	128	256	(3, 3, 3)	3	4
FEM	256	512	(3, 3, 3)	2	4

w stands for w_{AFFB} in AFFB, and w_{FEM} in FEM

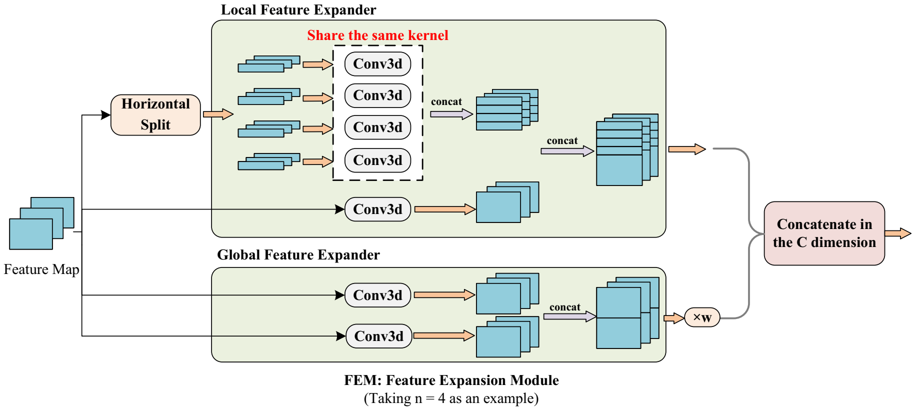


Fig. 4 The illustration of FEM. Local feature expander: we use a similar structure as AFFB, where the results of the two convolutional parts are concatenated in the height dimension as the local features. Global feature expander consists of two independent 3D convolutions, the same as local feature expander, and the result of the convolution is concatenated in the height dimension as the global feature. The output of local feature expander and global feature expander are concatenated in the channel dimension

focal convolution layer with the output of the common convolution layer, thereby enriching the final gait features and making them more representative.

The structure of the FEM is shown in Fig. 4. The FEM consists of two parts, the first is a local feature expander, and the second is a global feature expander. The global feature expander uses two parallel 3D convolutions to process the input, and the processed results are stitched on the H dimension. Finally, the output of the global feature expander is multiplied by the trainable weight w_{FEM} and stitched to the output of the local feature expander in the channel dimension to form the gait features. Assume that the output of the adaptive feature extractor is $V_{whole} \in C_{in} \times S \times H \times W$, where C_{in} is the number of input channels to the FEM, S is the length of the feature map sequence, and (H, W) is the size of each frame of the feature map. Similar to the analysis of the local feature extraction block in AFFB in Section 3.2, we divide each frame of the input feature map level into n parts, denoted as $V_{part} = \{V_{part}^{i} \mid i = 1, 2, ...n\} \in \mathbb{R}^{C_{in} \times S \times \frac{H}{n} \times W}$. We set the 3D convolution in the local feature expander as $f_{LE1}^{3\times3\times3}(\cdot)$ and $f_{LE2}^{3\times3\times3}(\cdot)$, and the 3D convolution in the global feature expander as $f_{GE1}^{3\times3\times3}(\cdot)$ and $f_{GE2}^{3\times3\times3}(\cdot)$, then the output of the global feature expander can be expressed as:

$$Y_{GE} = cat \left\{ \begin{array}{c} f_{GE1}^{3\times3\times3}(V_{whole}) \\ f_{GE2}^{3\times3\times3}(V_{whole}) \end{array} \right\} \in \mathbb{R}^{C_{out} \times S \times 2H \times W}, \tag{4}$$

and the output of local feature expander can be expressed as:

$$Y_{LE1} = cat \left\{ \begin{array}{c} f_{LE1}^{3\times3\times3}(V_{part}^{1}) \\ f_{LE1}^{3\times3\times3}(V_{part}^{2}) \\ \cdots \\ f_{LE1}^{3\times3\times3}(V_{part}^{n}) \end{array} \right\} \in \mathbb{R}^{C_{out} \times S \times H \times W}, \tag{5}$$

$$Y_{LE2} = f_{LE2}^{3\times3\times3}(V_{whole}) \in \mathbb{R}^{C_{out} \times S \times H \times W}, \tag{6}$$

$$Y_{LE} = cat \left\{ \begin{array}{c} Y_{LE1} \\ Y_{LE2} \end{array} \right\} \in \mathbb{R}^{C_{out} \times S \times 2H \times W}, \tag{7}$$

where C_{out} is the number of output channels of the 3D convolution.

Based on the above expressions for global feature expander and local feature expander, the output of FEM can be expressed as:

$$Y_{FEM} = cat\left\{\begin{matrix} w_{FEM} \cdot Y_{GE} \\ Y_{LE} \end{matrix}\right\} \in \mathbb{R}^{2C_{out} \times S \times 2H \times W}. \tag{8}$$

The output of the FEM is operated by GeM [13] and fully connected layers to obtain the final gait features. The specific parameters of FEM in the model are shown in Table 1.

3.4 Optimization

To optimize AdaptiveGait, the objective composed of triplet loss and cross-entropy loss. The triplet loss enhances the model's ability to distinguish between different individuals' gait features by minimizing intra-class differences and maximizing inter-class differences. Meanwhile, the cross-entropy loss optimizes classification accuracy, ensuring that the model accurately labels each gait pattern. The loss function is represented as follows:

$$L = L_{tri} + L_{ce}. \tag{9}$$

4 Experiments

We used two publicly available gait databases for performance testing of the model: CASIA-B [31] and OU-MVLP [21]. In this section, the specific conditions of the two databases are introduced and compared with other advanced gait recognition methods. Finally, based on the dataset CASIA-B, a detailed ablation study of the model is conducted to verify the function of the proposed module.

4.1 Datasets

CASIA-B The CASIA-B [31] dataset is the most commonly used cross-view gait database in gait recognition. It contains 124 subjects, each corresponding to ten gait sequences, of which six groups are tested under normal walking conditions, numbered by NM#01-NM#06. Two groups are tested under conditions of carrying a bag, numbered by BG#01-BG#02. The last two groups are sampled under conditions of wearing a coat, numbered by CL#01-CL#02. Each gait sequence contains 11 different sampling angles (ranging from 0° to 180°, with a sampling interval of 18°). In the training phase, we use 74 subjects to train the model based on the LT [25] experimental setup, and the remaining 50 subjects are used for testing. In the testing phase, the sequences NM#01-NM#04 are taken as the gallery set, and the sequences NM#05-NM#06, CL#01-CL#02 and BG#01-BG#02 are treated as the probe set to evaluate our model performance.

OU-MVLP The OU-MVLP [21] dataset is one of the most extensive public gait datasets. The dataset contains gait video sequences of 10307 subjects aged 2-87 years. Each subject corresponds to two sets of sequences, numbered by Seq#00 and Seq#01. Each sequence contains 14 different sampling angles (0°-90° and 180°-270°, with a sampling interval of 15°). Since the OU-MVLP dataset contains more subjects, it can better evaluate the model's generalization potential. In the training phase, this paper uses 5,153 subjects as training data based on the experimental setup in [2], and the remaining 5,154 subjects are used for testing.

During the testing phase, seq#01 is used as the gallery set, and seq#00 is taken as the probe set to evaluate the model's performance.

4.2 Implementation details

We used the method mentioned in [2] to preprocess the CASIA-B and OUMVLP datasets, and obtained a gait silhouette sequence normalized to 64×44. Next, we inputted the preprocessed gait silhouette sequence into the network for training. The specific training parameters of the network are shown in Table 1. We used Adam as the optimizer [9] and set the hyperparameter in triplet loss to 0.2. The batch sizes were set to (8, 8) and (16, 8) in the datasets CASIA-B and OUMVLP, respectively. For the CASIA-B dataset, Adam's weight attenuation was set to 5e-4. The learning rate was initially set to 1e-4 and reset to 1e-5 after an iteration of 70K rounds. We used an NVIDIA 4090 GPU training model and iterated 80K rounds. During the training of OUMVLP, the parameters w_{AFFB} and w_{FEM} in (3) and (8) were set to 11 and 6. For the OUMVLP dataset, the weight attenuation of Adam was initially set to 0 and reset to 5e-4 after 200K rounds, and the learning rate was initially set to 1e-4 and reset to 1e-5 and 5e-6 after 150K and 200K rounds. We used two NVIDIA 4090 GPU training models and iterated 270K rounds. During the training of OUMVLP, the parameters w_{AFFB} and w_{FEM} in (3) and (8) were set to 3 and 2.

4.3 Comparison with typical methods

Evaluation on CASIA-B [31] To demonstrate the effectiveness of the proposed method, we compared it with more typical models in the field of gait recognition. These include GaitSet [2], which is based on global features; GaitPart [3], which focuses mainly on local features; and GaitGL [13], which combines global and local features. Other methods include GLN [5], ESNet [6], and LagrangeGait [1]. The rank-1 accuracy of our method on CASIA-B for different walking patterns is shown in Table 2. Our method achieves a significant improvement under different walking conditions. The accuracies of our method in NM, BG and CL conditions are 97.8%, 95.1% and 86.0%, respectively. Compared with the baseline GaitGL, it is 0.4%, 0.6%, and 2.4% higher in each index, respectively. And compared to other typical gait recognition models, we obtained leading accuracy in all three different gait conditions. In addition, the average rank-1 accuracy of our method on CASIA-B is 93.0%. In contrast, the average rank-1 accuracy of LagrangeGait was 92.4%. The average accuracy of our method has surpassed the advanced gait recognition method LagrangeGait.

Evaluation on OU-MVLP [21] To demonstrate the generalizability of the proposed method, we compared the OU-MVLP with some typical models in the field of gait recognition. Since OU-MVLP contains more subjects and perspectives, it is an excellent test for the model's generalizability. The Rank-1 accuracies on OU-MVLP are shown in Table 3. Our method achieves the optimum in the 45°, 180°, and 240° perspectives, and the accuracy of our model is 89.9%, which is very close to the 90.0% result of the LagrangeGait [1]. The reason for this phenomenon is that the LagrangeGait uses additional motion extraction branches and visual embedding branches along with gait feature extraction, which is more expensive to train compared to our model. Compared to the baseline GaitGL, the average accuracy of our model improves by 0.2%, demonstrating the better generalization of our model.

Table 2 Rank-1 accuracy (%) on the CASIA-B [31] dataset

Gallery NM#1-4		0°-180°											
Prode		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	Mean
NM#5-6	GaitSet [2]	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	GaitPart [3]	94.1	98.6	99.3	98.5	94.0	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GLN [5]	93.2	**99.3**	**99.5**	**98.7**	96.1	95.6	97.2	98.1	**99.3**	98.6	90.1	96.9
	GaitGL[13]	96.0	98.3	99.0	97.9	96.9	95.4	97.0	98.9	**99.3**	98.8	94.0	97.4
	ESNet [6]	95.6	98.6	99.1	97.9	96.7	94.4	96.9	98.7	**99.3**	98.6	95.1	97.4
	LagrangeGait [1]	95.7	98.1	99.1	98.3	96.4	95.2	**97.5**	**99.0**	**99.3**	98.9	94.9	97.5
	Ours	**96.1**	98.2	99.1	97.9	**97.1**	**96.0**	97.3	**99.0**	**99.3**	**99.2**	**96.5**	**97.8**
BG#1-2	GaitSet [2]	83.8	91.2	91.8	88.8	83.3	81.0	**94.1**	90.0	92.2	94.4	79.0	87.2
	GaitPart [3]	89.1	94.8	96.7	95.1	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	GLN [5]	91.1	**97.7**	**97.8**	95.2	92.5	**91.2**	92.4	96.0	97.5	94.9	88.1	94.0
	GaitGL [13]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	98.2	96.9	91.5	94.5
	ESNet [6]	92.7	95.9	96.3	94.9	93.2	87.7	90.9	96.2	97.3	96.9	91.7	94.0
	LagrangeGait [1]	**94.2**	96.2	96.8	**95.8**	94.3	89.5	91.7	**96.8**	98.0	97.0	90.9	94.6
	Ours	93.8	96.3	97.2	95.5	**94.8**	90.4	93.2	96.7	**98.5**	**97.3**	**92.6**	**95.1**
CL#1-2	GaitSet [2]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart [3]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GLN [5]	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	GaitGL [13]	76.6	90.0	90.3	87.1	84.5	79.0	84.1	87.0	87.3	84.4	69.5	83.6
	ESNet [6]	75.6	89.2	92.4	**90.3**	84.3	80.2	83.0	86.3	89.0	83.9	69.8	84.0
	LagrangeGait [1]	77.4	90.6	93.2	90.2	84.7	80.3	**85.2**	87.7	89.3	**86.6**	71.0	85.1
	Ours	**79.1**	**91.8**	**93.9**	90.1	**85.1**	**81.4**	**85.2**	**88.9**	**90.3**	86.4	**74.1**	**86.0**

Results for 11 views and different walking conditions are included. Excluding the case of the same view

The bolded data in the table represent the optimal results under the respective experimental conditions, facilitating readers to quickly grasp the key points

Table 3 Rank-1 accuracy (%) on the OU-MVLP [21] dataset, excluding the case of the same view

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GaitSet [2]	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
GaitPart [3]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [5]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	**90.5**	**90.6**	89.6	89.3	88.5	89.2
GaitGL [13]	84.9	90.2	91.1	**91.5**	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
ESNet [6]	84.8	89.6	91.0	91.3	90.7	90.4	89.9	88.5	87.5	90.1	90.2	89.4	89.3	88.5	89.4
LagrangeGait [1]	**85.9**	**90.6**	**91.3**	**91.5**	**91.2**	**91.0**	**90.6**	88.9	**89.2**	**90.5**	**90.6**	**89.9**	**89.8**	**89.2**	**90.0**
Ours	85.5	90.2	91.2	**91.5**	91.1	90.9	90.5	**89.1**	88.6	90.3	90.4	**89.9**	89.7	89.0	89.9

The bolded data in the table represent the optimal results under the respective experimental conditions, facilitating readers to quickly grasp the key points

Table 4 Ablation experiments performed on AFFB and FEM under the CASIA-B dataset.(Rank-1, %)

Configurations	NM	BG	CL	Mean
Baseline(GaitGL)	97.4	94.5	83.6	91.8
+AFFB	97.5	94.5	84.5	92.1
+FEM	97.5	94.9	84.8	92.4
+AFFB+FEM	**97.8**	**95.1**	**86.0**	**93.0**

The bolded data in the table represent the optimal results under the respective experimental conditions, facilitating readers to quickly grasp the key points

4.4 Ablation study

Analysis of AFFB and FEM We conducted ablation experiments on the modules to verify the effectiveness of AFFB and FEM in gait feature extraction and the necessity of adaptivity. Since CAISA-B [31] contains more walking conditions to validate the modules in different situations, we conducted ablation experiments on this dataset.

The results of the ablation experiments are shown in Table 4. With GaitGL [13] selected as Baseline and with AFFB alone, NM increased by 0.1%, CL increased by 0.9%, BG did not change, and the average accuracy improved by 0.3%. The results show that AFFB has superior extraction ability for gait features. In addition, with FEM alone, results showed a 0.1% rise in NM, a 0.4% rise in BG, a 1.2% rise in CL, and a 0.6% increase in average accuracy over GaitGL. This demonstrates the crucial role of the FEM in the adequate representation of gait features. Using the two together, we concluded with a 0.4% increase in NM, a 0.6% increase in BG, a 2.4% increase in CL, and a 1.2% increase in average accuracy. This is a significant improvement on the advanced model, and demonstrates the effectiveness of our proposed method.

Analysis of different feature extractors We conducted ablation experiments using four feature extractors, 3D convolution, FConv [3], GLFE [13] and AFFB, to verify that AFFB can extract the complete gait features better compared to other methods. We conducted the experiments on CASIA-B [31] with GaitGL [13] as the baseline. The results are shown in Table 5. It can be seen that it makes sense to extract the AFFB module compared to other feature extraction methods. AFFB adaptively balances global features with local features and achieves better results on NM, BG, and CL compared to other feature extractors.

Analysis of FEM structure We further analyzed the structure of FEM. The specific results are shown in Table 6. The LFE and GFE in the table represent the Local Feature Expander and Global Feature Expander of FEM, respectively. The results show that the recognition accuracy is low when experiments are conducted using only LFE and GFE. This

Table 5 Ablation experiments using 3D convolution, FConv, GLFE, and AFFB for feature extraction under the CASIA-B dataset.(Rank-1, %)

Configurations	NM	BG	CL	Mean
Baseline + 3D Conv	96.9	93.9	83.5	91.4
+ FConv	96.6	93.4	82.5	90.8
+ GLFE	97.4	**94.5**	83.6	91.8
+ AFFB	**97.5**	**94.5**	**84.5**	**92.1**

The bolded data in the table represent the optimal results under the respective experimental conditions, facilitating readers to quickly grasp the key points

Table 6 Ablation experiments for the components of FEM using the CASIA-B dataset

Method	Fusion	NM	BG	CL	Mean
Analysis of LFE and GFE					
LFE	-	97.3	94.6	84.1	92.0
GFE		97.3	94.6	84.8	92.2
Analysis of fusion methods					
LFE+GFE	cat in H	97.4	94.5	84.3	92.1
	cat in C	**97.8**	**95.1**	**86.0**	**93.0**

"cat in H" denotes "concatenate in the height dimension of feature maps", while "cat in C" indicates "concatenate in the channel dimension of feature maps".(Rank-1, %)
The bolded data in the table represent the optimal results under the respective experimental conditions, facilitating readers to quickly grasp the key points

is because the information extracted by LFE and GFE is relatively singular: LFE is more advantageous in extracting fine-grained features due to the horizontal division of the feature maps, while GFE can better preserve the complete spatio-temporal features using 3D CNN. For the feature fusion problem of LFE and GFE, we tested two methods: concatenate in height dimension and concatenate in channel dimension.The results demonstrate that the highest recognition accuracy is achieved when concatenating in the channel dimension. This can be attributed to the increase in the number of neurons in the subsequent fully connected layer after concatenation in the channel dimension. This expansion of gait features along the channel dimension allows for a more effective representation of gait to be learned.

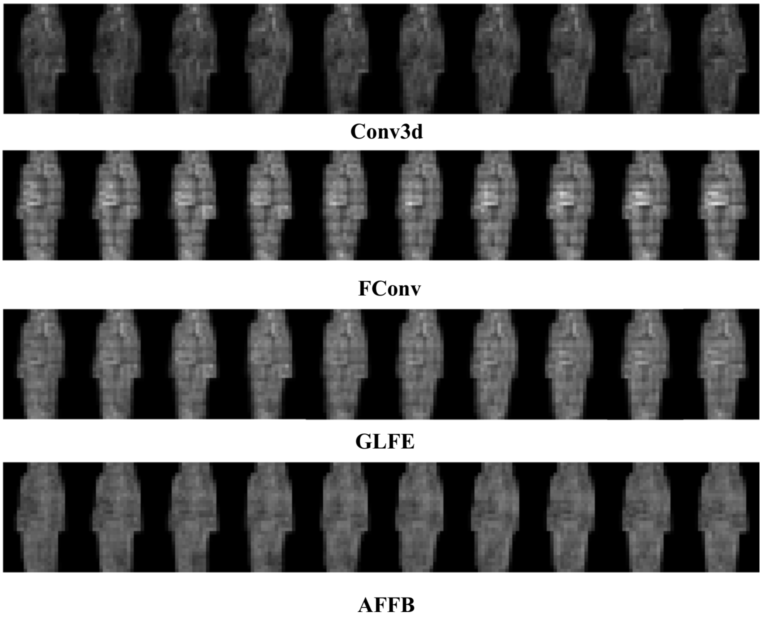
4.5 Complexity analysis

We performed a complexity analysis of our model as shown in Table 7. We conducted separate experiments on the CASIA-B dataset using the same settings on the same device for our model and GaitGL [13] models. Since LagrangeGait [1] does not currently have a standardized code, no comparison was made with it. From the results, it can be seen that our model has a higher number of parameters compared to GaitGL[13]. This indicates that our model is more powerful in terms of representation capability, but it also means that it requires more computational resources during training and inference. As can be seen from the table, on the CASIA-B dataset, our model has an inference speed of 165.07it/s, while GaitGL [13] has an inference speed of 238.62it/s. Despite exhibiting slightly slower inference speed compared to GaitGL [13], our model maintains an acceptable rate, aligning well with the inference speed standards prevalent in the domain of gait recognition.

Table 7 Complexity results for GaitGL and our method (AdaptiveGait) under CASIA-B

Model	Parameters Count	Inference Speed
GaitGL	3.10M	238.62it/s
Ours	7.73M	165.07it/s

The results for both models were obtained by computing under two NVIDIA GeForce RTX 4090



Conv3d

FConv

GLFE

AFFB

Fig. 5 Visualization of the gait feature maps extracted using the different methods. The first row shows the feature maps extracted by common 3D convolution. The second row shows the feature maps extracted using the 3D convolution of the FConv [3] idea. The third row shows the feature map extracted by the GLFE in GaitGL [13]. The fourth row shows the feature map extracted by AFFB

4.6 Visualization

Figure 5 shows the gait feature maps under different feature extraction methods. The first row shows the gait feature map extracted by normal convolution. It can be seen that the normal feature map information is more comprehensive, but insufficient attention is given to the body details, and some information is very fuzzy. The second row shows the gait feature map extracted using the FConv [3] idea, and it can be seen that the feature map appears to be significantly chunked. Although detailed information of different parts of the body is more effectively extracted, the connection between these parts is weakened, and the features at the dividing line of different parts are lost. The third row shows the gait feature map extracted from GaitGL [13]. Although the fusion of global and local features has been considered, chunking is still apparent, and the gait features are not fully expressed. The last row shows the features extracted using the adaptive feature extractor. It can be seen that the features at the dividing line are significantly supplemented, and the features of each body part are enhanced. Finally, a more complete gait feature extraction is obtained.

5 Conclusion

This paper proposes the application of an adaptive feature fusion block for gait recognition, and existing methods of gait feature extraction were shown to be improved using an adaptive feature fusion technique. Using visualization methods, we successfully compensated for feature map loss due to horizontal segmentation during local feature extraction. In addition, we also propose a feature expansion module to strengthen the connection of body parts in the feature map by introducing more global information, while enriching the temporal information in the feature space. We demonstrate the method's feasibility on two datasets commonly used for gait recognition. Our proposed method can extract complete gait features more effectively, and we hope to provide some ideas for additional gait recognition work in the future. Our concept is also adaptable to pedestrian re-recognition and other recognition fields.

Acknowledgements This work was supported in part by the Key R&D Project of Shandong Province (2022CXGC010503), in part by 2022 Industry Leading Talents Support Program "Hai You Project" (Research and industrialization of management and control system for information technology service industry based on AI and edge computing), and in part by the National Natural Science Foundation of China (61876099).

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

Ethical standard Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

1. Chai T, Li A, Zhang S et al (2022) Lagrange motion analysis and view embeddings for improved gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 20249–20258
2. Chao H, He Y, Zhang J et al (2019) Gaitset: Regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI conference on artificial intelligence, pp 8126–8133
3. Fan C, Peng Y, Cao C et al (2020) Gaitpart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14225–14233
4. Ghiasi G, Lin TY, Le QV (2019) Nas-fpn: learning scalable feature pyramid architecture for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 7036–7045
5. Hou S, Cao C, Liu X et al (2020) Gait lateral network: Learning discriminative and compact representations for gait recognition. In: Part IX (ed) Paper presented at the computer vision-ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings. Springer, pp 382–398
6. Huang T, Ben X, Gong C et al (2022) Enhanced spatial-temporal salience for cross-view gait recognition. IEEE Trans Circuits Syst Video Technol 32(10):6967–6980
7. Jiao L, Xie C, Chen P et al (2022) Adaptive feature fusion pyramid network for multi-classes agricultural pest detection. Comput Electron Agricul 195:106827
8. Ju M, Luo J, Wang Z et al (2021) Adaptive feature fusion with attention mechanism for multi-scale target detection. Neural Comput Appl 33:2769–2781
9. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
10. Li X, Makihara Y, Xu C et al (2020) End-to-end model-based gait recognition. In: Proceedings of the Asian conference on computer vision
11. Li Y, Yao H, Duan L et al (2019) Adaptive feature fusion via graph neural network for person re-identification. In: Proceedings of the 27th ACM international conference on multimedia, pp 2115–2123
12. Liao R, Yu S, An W et al (2020) A model-based gait recognition method with body pose and human prior knowledge. Pattern Recogn 98:107069
13. Lin B, Zhang S, Yu X (2021) Gait recognition via effective global-local feature representation and local temporal aggregation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 14648–14656
14. Liu A, Yang Y, Sun Q et al (2018a) A deep fully convolution neural network for semantic segmentation based on adaptive feature fusion. Paper presented at the 2018 5th international conference on information science and control engineering (ICISCE), IEEE, pp 16–20

15. Liu S, Qi L, Qin H et al (2018b) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768
16. Luo J, Cui W, Xu S et al (2023) A dual-branch spatio-temporal-spectral transformer feature fusion network for eeg-based visual recognition. IEEE Trans Industr Inform
17. Qiao D, Zulkernine F (2023) Adaptive feature fusion for cooperative perception using lidar point clouds. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1186–1195
18. Qin H, Chen Z, Guo Q et al (2021) Rpnet: Gait recognition with relationships between each body-parts. IEEE Trans Circuits Syst Video Technol 32(5):2990–3000
19. Shang R, Zhang J, Jiao L et al (2020) Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images. Remote Sens 12(5):872
20. Shiraga K, Makihara Y, Muramatsu D et al (2016) Geinet: View-invariant gait recognition using a convolutional neural network. Paper presented at the 2016 international conference on biometrics (ICB), IEEE, pp 1–8
21. Takemura N, Makihara Y, Muramatsu D et al (2018) Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans Comput Vis Appl 10:1–14
22. Tan M, Pang R, Le QV (2020) Efficientdet: Scalable and efficient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10781–10790
23. Teepe T, Khan A, Gilg J et al (2021) Gaitgraph: Graph convolutional network for skeleton-based gait recognition. Paper presented at the 2021 IEEE international conference on image processing (ICIP), IEEE, pp 2314–2318
24. Wang M, Zhao L, Yue Y (2023) Pa3dnet: 3-d vehicle detection with pseudo shape segmentation and adaptive camera-lidar fusion. IEEE Trans Industr Inform
25. Wu Z, Huang Y, Wang L et al (2016) A comprehensive study on cross-view gait based human identification with deep cnns. IEEE transactions on pattern analysis and machine intelligence 39(2):209–226
26. Xu C, Makihara Y, Li X et al (2020) Cross-view gait recognition using pairwise spatial transformer networks. IEEE Trans Circuits Syst Video Technol 31(1):260–274
27. Xu C, Liu H, Guan Z, Wu X, Tan J, Ling B (2022) Adversarial incomplete multiview subspace clustering networks. IEEE Trans Cybern 52(10):10490–10503. https://doi.org/10.1109/TCYB.2021.3062830
28. Xu C, Zhao W, Zhao J, Guan Z, Song X, Li J (2023) Uncertainty-aware multiview deep learning for internet of things applications. IEEE Trans Industr Inform 19(2):1456–1466. https://doi.org/10.1109/TII.2022.3206343
29. Xu C, Zhao W, Zhao J et al (2023) Progressive deep multi-view comprehensive representation learning. In: Proceedings of the AAAI conference on artificial intelligence, pp 10557–10565
30. Yang HH, Huang KC, Chen WT (2021) Laffnet: A lightweight adaptive feature fusion network for underwater image enhancement. Paper presented at the 2021 IEEE international conference on robotics and automation (ICRA), IEEE, pp 685–692
31. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. Paper presented at the 18th international conference on pattern recognition (ICPR'06), IEEE, pp 441–444
32. Zhao Q, Sheng T, Wang Y et al (2019) M2det: A single-shot object detector based on multi-level feature pyramid network. In: Proceedings of the AAAI conference on artificial intelligence, pp 9259–9266

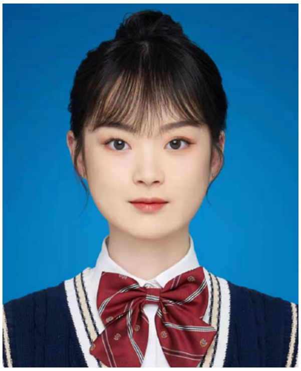
Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Tian Liang was born in Shandong, China, in 1999. He received the B.E. degree in School of Control Science and Engineering, Shandong University, Jinan, China, in 2021. He is currently pursuing the M.E. degree in the School of Control Science and Engineering, Shandong University, Jinan, China. His research interests include computer vision, deep learning, and gait recognition.

Zhenxue Chen was born in Shandong, China, in 1977. He received the B.S. degree in automatic from School of Electrical Engineering and Automation at Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from School of Information Science and Engineering at Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Image Recognition and Artificial Intelligence at Huazhong University of Science and Technology, Wuhan, China, in 2007. From 2012 to 2013, he was a visiting scholar with the Michigan State University, East Lansing, Michigan, USA. He is currently a professor with the School of Control Science and Engineering, Shandong University. His main areas of interest include image processing, pattern recognition, and computer vision, with applications to face recognition. He has published over 100 papers in refereed international leading journals/conferences such as IEEE T-II, IEEE T-CSVT, IEEE T-IFS, IEEE T-VT, IEEE T-ITS, Information Sciences, Neurocomputing, Neural Computing and Applications, and SP-IC, etc.

Chengyun Liu was born in Henan, China, in 1975. She received the B.S. degree in communication from Huazhong Normal University, Wuhan, China, in 1999, the M.S. degree in pattern recognition and intelligent systems from the Wuhan University of Science and Technology, Wuhan, in 2005, and the Ph.D. degree in pattern recognition and intelligent systems from Shandong University, Jinan, China, in 2016. She is currently an Associate Professor with the School of Control Science and Engineering, Shandong University. Her research interests include automatic target detection and recognition, image processing, and computer vision.

Jiyang Chen was born in Shandong China in 1991. He received the B.S. and M.S. degrees from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2013 and in 2017. He is currently pursuing the Ph.D. degree in Institute of Marine Science and Technology, Shandong University, Qingdao, China. His research interests include machine learning, deep learning, and face recognition, etc.



Yuchen Hu was born in Shandong, China, in 2000. She received the B.E. degree in School of Control Science and Engineering, Shandong University, Jinan, China, in 2022. She is currently pursuing the M.E. degree in the School of Control Science and Engineering, Shandong University, Jinan, China. Her research interests include computer vision, deep learning, and gait recognition.

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

Authors and Affiliations

Tian Liang[1] · Zhenxue Chen[1] · Chengyun Liu[1] · Jiyang Chen[2,3] · Yuchen Hu[1] · Q. M. Jonathan Wu[4]

Tian Liang
202134867@mail.sdu.edu.cn

Jiyang Chen
chenjy@sdas.org

Yuchen Hu
202214803@mail.sdu.edu.cn

Q. M. Jonathan Wu
jwu@uwindsor.ca

[1] School of Control Science and Engineering, Shandong University, 250061 Jinan, China

[2] Institute of Marine Science and Technology, Shandong University, 266237 Qingdao, China

[3] Shandong Zhengzhong Information Technology CO., LTD., 250098 Jinan, China

[4] Department of Electrical and Computer Engineering, University of Windsor, ON N9B 3P4 Windsor, Canada