Neural Processing Letters (2024) 56:133
https://doi.org/10.1007/s11063-024-11554-8

Human Gait Recognition Based on Frontal-View Walking Sequences Using Multi-modal Feature Representations and Learning

Muqing Deng[1] · Zebang Zhong[1] · Yi Zou[1] · Yanjiao Wang[1] · Kaiwei Wang[2] · Junrong Liao[3]

Accepted: 11 February 2024

Abstract
Despite that much progress has been reported in gait recognition, most of these existing works adopt lateral-view parameters as gait features, which requires large area of data collection environment and limits the applications of gait recognition in real-world practice. In this paper, we adopt frontal-view walking sequences rather than lateral-view sequences and propose a new gait recognition method based on multi-modal feature representations and learning. Specifically, we characterize walking sequences with two different kinds of frontal-view gait features representations, including holistic silhouette and dense optical flow. Pedestrian regions extraction is achieved by an improved YOLOv7 algorithm called Gait-YOLO algorithm to eliminate the effects of background interference. Multi-modal fusion module (MFM) is proposed to explore the intrinsic connections between silhouette and dense

Muqing Deng and Zebang Zhong have contributed equally to this work.

✉ Yanjiao Wang
yjwang12@gdut.edu.cn

✉ Kaiwei Wang
hanxingfly@126.com

Muqing Deng
mqdeng@gdut.edu.cn

Zebang Zhong
zhongzebang2020@163.com

Yi Zou
gdutzouyi@qq.com

Junrong Liao
liaojunrongcsg@dingtalk.com

1 School of Automation, Guangdong University of Technology, Guangzhou, China

2 Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology, China Electronic Product Reliability and Environmental Testing Research Institute, Guangzhou, China

3 Project Development Center, CSG Power Generation Energy Storace Technology Co., Ltd, Guangzhou, China

Published online: 02 April 2024

optical flow features by using squeeze and excitation operations at the channel and spatial levels. Gait feature encoder is further used to extract global walking characteristics, enabling efficient multi-modal information fusion. To validate the efficacy of the proposed method, we conduct experiments on CASIA-B and OUMVLP gait databases and compare performance of our proposed method with other existing state-of-the-art gait recognition methods.

Keywords Gait recognition · Multi-modal feature learning · Dense optical flow · Deep learning

1 Introduction

Gait recognition, a new kind of non-contact biometric technique, has received increasing attentions since the outbreak of COVID-19 [1]. In contrast to conventional biometric techniques, e.g. face or fingerprint recognition, gait recognition has great advantages of long-distance non-contact individual identification without the cooperations of subjects, making it particularly attractive for security-sensitive surveillance applications [2, 3].

Much progress has been reported in the area of gait recognition, however, most of these existing works adopt lateral-view parameters as gait features. These methods usually suffer from severe occlusion problems of human silhouette between left and right legs [4, 5]. In addition, to obtain walking sequences of the whole gait cycle, it is necessary to place cameras at a considerable distance from subjects, which requires large area of data collection environment and limits the applications of gait recognition in real-world practice. Attempts to resolve these dilemmas have resulted in the development of frontal-view gait recognition methods.

To name a few, Tahmoush et al. [6] presented a frontal-view gait recognition algorithm based on the data stream of remote micro-Doppler radar that enables individual detection, tracking and identification. Barnich et al. [7] extracted gait time-frame interval features from frontal-view walking sequences and built a real-time system for pedestrian identification using improved machine learning algorithms. Matovski et al. [8] combined the advantages of gait energy image (GEI) and gait entropy image (GEnI) and proposed a hybrid frontal-view gait recognition method. It can be observed that the aforementioned methods can sort out the frontal-view gait recognition problem to some extent. However, it is still challenging to deal with the changes of human silhouettes especially when complex backgrounds and external environmental factors are taken into consideration [9]. Additional depth cameras or radar sensors are involved in the existing frontal-view recognition works, leading to heavy hardware system burdens as well. Therefore, our proposed method tackle these issues as the core objective in this paper, and propose a new frontal-view gait recognition method based on multi-modal feature representations and learning.

In this study, we propose a new frontal-view gait recognition method using multi-modal feature representations and learning. First, we propose that, the extracted binary frontal-view gait silhouettes are characterized with two different kinds of frontal-view gait features representations: holistic silhouette and dense optical flow. Pedestrian regions extraction is achieved by an improved YOLOv7 algorithm to eliminate the effects of background interference. Holistic silhouettes are used to describe the spatial characteristic during human walking, while dense optical flow can be designed to reflect the temporal dynamics information of human walking. Second, multi-modal fusion module (MFM) is proposed to explore the intrinsic connections between silhouette and dense optical flow features by using squeeze

and excitation operations at the channel and spatial levels. Gait feature encoder is further used to extract global walking characteristics, enabling efficient multi-modal information fusion. The aforementioned two kinds of gait representations reflect the spatial and temporal changes of walking silhouettes in two different aspects, and are fused by using the proposed multi-modal fusion module for a better recognition performance.

The main contributions of this paper are as follows: (1) Our proposed method facilitates the application of gait-based individual identification in real-world scenarios by using only frontal-view walking sequences, particularly in security-sensitive narrow corridors. (2) Our proposed method combines holistic silhouette and dense optical flow to capture the spatial characteristic and temporal characteristics during the gait cycle to further improve the recognition rate. (3) Our proposed method innovatively designs a multi-modal fusion module, which utilizes squeeze and excitation operations to simulate the interaction between different modes. Additionally, we have developed a gait feature encoder, called Gait-Encoder, to globally encode the gait vector for efficient fusion of silhouette and dense optical flow features. (4) Our proposed method achieves promising results on publicly available CASIA-B and OUMVLP gait databases, and outperforms existing other gait recognition methods. Moreover, our method maintains reliable results even under different walking conditions.

2 Related Work

Existing gait recognition methods can be roughly classified into two categories: model-based and appearance-based. Appearance-based representations can be further classified into silhouette-based and pose-based methods [10, 11]. Walking silhouettes contain useful temporal information for gait recognition, however, they are highly sensitive to pedestrian attire or carrying statuses. Pose-based gait recognition methods typically use skeleton parameters for individual identification and usually suffer from the problems of occlusion and human body modeling.

In recent years, deep learning-based gait recognition methods have gained increasing attentions due to their superior performance. Neural network-based feature extraction techniques [12, 13] and graph-based gait representations [14–16] have been widely adopted. To name a few, Chao et al. [17] used frame-level features to construct gait templates and employed convolutional neural networks for silhouette-based feature extraction. Shiraga et al. [18] proposed a lightweight convolutional neural network (Geinet) for gait recognition. Tong et al. [19] introduced a restricted triple network to address the effect of viewpoint changes on gait recognition, and they also introduced triple loss to further improve the recognition rates of gait recognition.

In fact, gait characteristics underlying one single modal are limited and not enough to develop a high-accurate recognition system. Based on this assumption, one possible method is to develop multi-modal fusion methods [20] for gait recognition. Papavasileiou et al. [21] proposed a multi-modal learning method that utilized accelerometer and ground contact force data, and designed an automatic encoding network for feature extraction. Manuel et al. [22] proposed a robust gait recognition framework, UGaitNet, that integrates various types of input modalities including pixel gray values, depth maps, and silhouettes to obtain advanced gait descriptors. Liu et al. [23] proposed a lightweight dual-channel convolutional neural network, DC-DSCNN, which combines gait cycle sequence data and Gramian angular field (GAF) images to achieve gait recognition with higher real-time functionality for wearable devices. Li et al. [24] proposed a multi-modal gait recognition method that combines sil-

houette and human skeleton modalities to overcome the effects of pedestrian clothing and carrying. However, these methods only utilize simple stitching fusion for the feature maps of each modality at the multi-modal fusion level and do not fully consider the intrinsic connection between each modality. To address this issue, we propose a multi-modal fusion module (MFM) that models the interactions between different modalities and learns different motion patterns between each modality, enabling efficient gait recognition.

As for frontal-view gait recognition, several pioneering works have confirmed the feasibility of frontal-view gait parameters for the task of individual identification. Barnich et al. [7] fit gait features within closed silhouettes from a set of rectangles to frontal-view walking sequences. However, this method was limited in its ability to accurately predict covariate factors when walking appearance changes, as the size of the rectangles changed accordingly. Goffredo et al. [25] proposed an algorithm using 3D gait volumes for frontal view gait recognition, but drop of recognition rates were taken place when clothing and carrying conditions changed. Soriano et al. [5] utilized Freeman Code for silhouette extraction in frontal-view gait recognition, but this method was highly sensitive to complex background environment. After these pioneering works, a vast literature has accumulated on frontal-view gait recognition. Combined use of depth and RGB data streams from Kinect sensors was further investigated and represented as pose depth volume for frontal-view gait recognition [26]. Depth parameters from time-of-flight cameras were extracted and a four-part method was presented in [27]. Spatial-temporal characteristics underlying depth data streams were calculated and used as the input of different deep learning models for frontal-view gait recognition [28]. Skeleton joint parameters from the depth data of popular Kinect sensor was extracted and used to develop a key-pose-based frontal-view gait recognition method [29]. Liao et al. [30] further adopted the OpenPose tool and 18-joint model for extraction of human joint angles as well as their joint movements. Joint features and bone features were calculated using skeleton data and a dual graph CNN model was used as classifier in [31].

3 Proposed Method

As shown in Fig. 1, our proposed method can be summarized in this section in the following three parts: multi-modal feature representation, target pedestrian areas detection and multi-modal fusion module. We extract holistic silhouette and dense optical flow as two different kinds of gait representations. Target pedestrian regions are extracted using an improved YOLOv7 algorithm for frontal-view gait sequences, and gait in-depth features can then be captured using convolutional neural network. Based on the captured gait features, we design a multi-modal fusion module to explore the intrinsic connections between the silhouette and dense optical flow features through squeezing and excitation operations. A gait feature encoder, Gait-Encoder, is further used in MFM to globally encode the gait feature vectors, thereby recognizing human gait through the encoded high-level semantic motion features.

3.1 Holistic Silhouette Extraction

We firstly extract holistic silhouette of each frontal-view frame to describe the spatial change characteristics during human walking. Original walking frames are converted into grayscale images to reduce computational complexity. Gaussian filtering and histogram equalization are performed to enhance contrast, and background subtraction is adopted to remove background noise. For the case of noise and isolated pixels in the extracted silhouette images, we apply

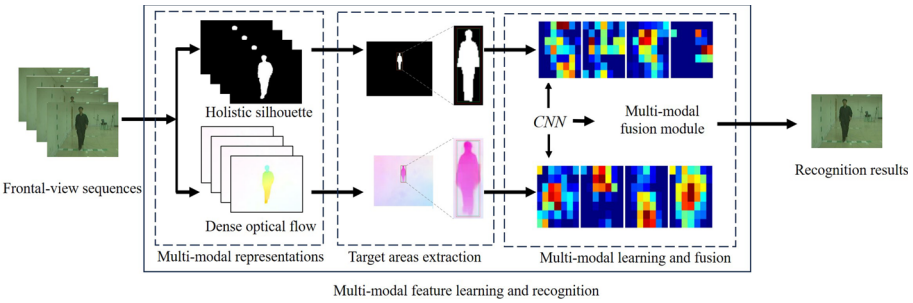

Fig. 1 The overall block diagram of the proposed method

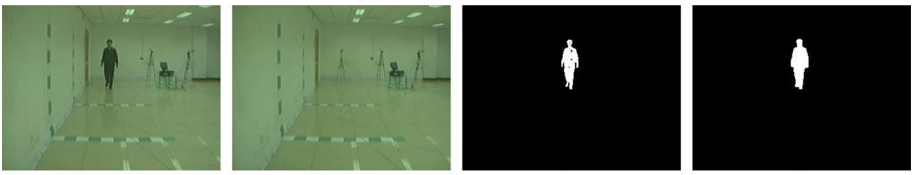

Fig. 2 Holistic silhouette extraction: **a** original frontal-view walking frame, **b** background image, **c** binary silhouette, **d** clean holistic silhouette

morphological processing to fill in holes and ensure that the gait silhouette has complete connectivity. The final clean gait silhouettes for normal walking are shown in Fig. 2.

3.2 Dense Optical Flow Extraction

Optical flow has been proved as an effective representation for video understanding, which presents instantaneous movement pattern of each pixel in image plane when the target object is moving. This section mainly extracts dense optical flow features of each frontal-view frame to describe the temporal change characteristics during human walking. We use point-to-point matching between two adjacent frames, and go through every pixel point to form dense optical flow for gait recognition.

Denote arbitrary two adjacent frames in the walking sequence as $I(x, y, t)$ and $I(x + u, y + v, t + 1)$, in which $I(\cdot)$ represents gray value calculation function, (x, y) represents pixel coordinates, t represents time, u and v represent pixel displacement at time t and $t + 1$, respectively. Based on constant brightness hypothesis, gray values of pixels at the same position in consecutive two images should be equal, that is, $I(x, y, t) = I(x+u, y+v, t+1)$. Using Taylor's expansion, gray value at the pixel point $(x + u, y + v, t + 1)$ can be obtained by first order approximation:

$$I(x + u, y + v, t + 1) \approx I(x, y, t) + u\Delta_x + v\Delta_y + \Delta_t \tag{1}$$

where Δ_x, Δ_y, Δ_t represent gradient of $I(x, y, t)$ in the x, y, and t directions, respectively. According to spatial consistency hypothesis, displacement between adjacent pixels should be continuous,u and v can be approximated as displacement difference between adjacent pixels, that is, $u = \delta x$ and $v = \delta y$. Therefore, Eq. 3 can be rewritten as:

Fig. 3 Dense optical flow extraction: **a** original frontal-view walking frame, **b** optical flow calculation, **c** visualization of optical flow using color coding, **d** smoothed optical flow images

$$I(x+u, y+v, t+1) \approx I(x, y, t) + \delta x \Delta_x + \delta y \Delta_y + \Delta_t \tag{2}$$

This is an equation for δ_x and δ_y, and we can solve it by minimizing the reprojection error to get the displacement vector (u, v) for every pixel point. After obtaining the extracted optical flow vector image, we use median filtering to smooth the original optical flow vector image to remove unnecessary noise and improve the accuracy and stability of optical flow vector image. Color coding is further used to visualize the optical flow vector image to form a new kind of gait representation. To improve generalization ability of the extracted dense optical flow image of gait, we overlay dense optical flow images of adjacent three frames $(i-1, i, i+1)$ by using weight ratio of (a, b, c). In our experiments, weight parameters (a, b, c) are set to (0.2, 0.6, 0.2). Figure 3 shows the extraction process of dense optical flow images for frontal-view walking sequence.

3.3 Target Pedestrian Areas Detection

The most significant distinguishing factor of frontal-view gait recognition is that height and width of pedestrian target region vary with distance of pedestrian to camera. To solve this problem, this section introduces a new method based on popular YOLOv7 algorithm for detecting pedestrians in holistic silhouettes and dense optical flow images.

3.3.1 Small Object Detection Layer

Since conventional YOLOv7 algorithm suffers from the problem of detecting small pedestrian areas, this paper proposes an improved YOLOv7-based algorithm, called Gait-YOLO, by introducing a small object detection layer for high-precision pedestrian detection.

The YOLOv7 includes three feature maps at scales of 80×80, 40×40, and 20×20, respectively. One of the most significant difference between frontal-view gait recognition in this paper and lateral-view gait recognition is the fact that pedestrians walk closer from a distance, the size of the pedestrian region varies from small to large. From the view of detecting large objects with a small receptive field and small objects with a large receptive field, this section maintains the unchanged scale of the baseline network output feature maps, and further introduces a detection layer with a receptive field of 160×160 in the detection head. This operation partitions the input image into 160×160 grid cells, enabling improved regression of prior boxes and adjustment for the detection of small target objects. The additional small target detection layer is highlighted in red boxes in Fig. 4.

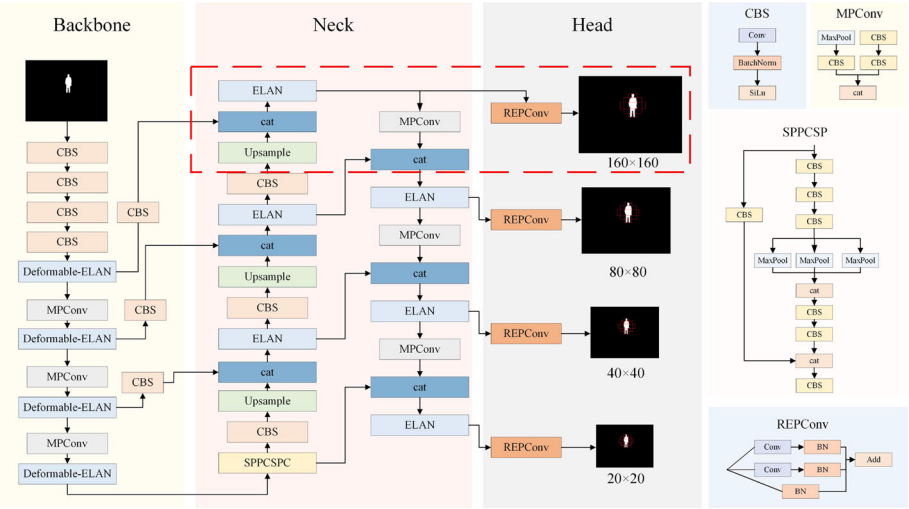

Fig. 4 Overall structure of Gait-YOLO model. (Color figure online)

3.3.2 Deformable Convolution V3

Conventional convolution operations are generally regular convolutions for sampling of fixed-sized regions. For frontal-view gait recognition, as pedestrian region continuously changes during walking movement, conventional convolution is not suitable for the regression task of pedestrian movement region. Deformable convolution operation, in this sense, can help to adjust the receptive field according to movement and variations in pedestrian's appearance. Therefore, a deformable-ELAN module based on deformable convolution v3 is designed in this section to address the limitations of conventional convolutional layers in capturing long-range dependencies and achieve global representation of pedestrian gait:

$$y(p_0) = \sum_{g=1}^{G} \sum_{k=1}^{K} w_g m_{gk} x_g (p_0 + p_k + \Delta p_{gk}) \tag{3}$$

where p_0 represents the current pixel point, G is the number of aggregation groups, K is the total number of sample points. For the g-th group, $w_g \in R^{C \times C'}$ represents the position-independent projected weight of a group, with the dimension of the group $C' = C/G$. $m_{gk} \in \mathbb{R}$ represents modulation scalar of the k-th sampling point in the g-th group. $x_g \in \mathbb{R}^{C' \times H \times W}$ is the sliced input feature vector. Δp_{gk} represents the corresponding offsets of p_k in the g-th group. The proposed Deformable-ELAN module can not only increase the expressiveness of the operator but also share convolution weights, reducing algorithm complexity and enhancing training stability. The structures of the ELAN module and Deformable-ELAN module are illustrated in Fig. 5.

The structure of the proposed Gait-YOLO algorithm including ELAN module and Deformable-ELAN module are shown in Fig. 4. The Gait-YOLO model is used to extract regions of interest in the holistic silhouettes and dense optical flow images, as shown in Fig. 6. The extracted pedestrian regions of silhouette images and dense optical flow images are resized into 64 × 44 images and used as the dual-channel input for final recognition

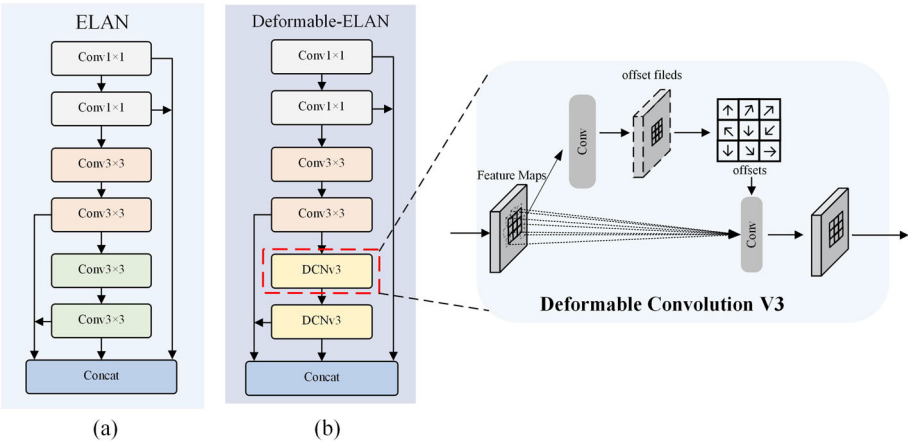

Fig. 5 The structures of **a** ELAN and **b** deformable-ELAN modules

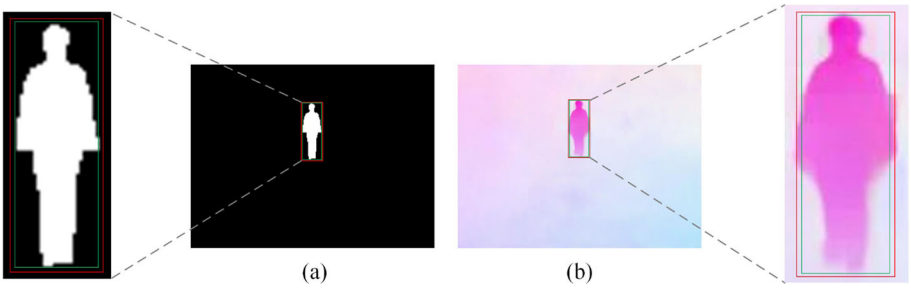

Fig. 6 Results of pedestrian detection by Gait-YOLO: **a** silhouette, **b** dense optical flow. The red box is the ground-truth box and the green one is the detecting box

task. The proposed Gait-YOLO algorithm can achieve extraction of pedestrian areas from both silhouette and dense optical flow by adopting a small target detection layer and a Deformable-ELAN module. Despite that conventional image processing methods do well in target pedestrian areas detection of binary silhouettes, they fail to obtain satisfactory performance in images with rich color changes, such as dense optical flow graphs. Additionally, the proposed Gait-YOLO algorithm can obtain promising performance in pedestrian area detection of binary silhouettes, dense optical flows and even original RGB frame sequences only with small amount of pre-training, which outperforms conventional image processing methods.

It is also noted that the baseline YOLOv7 model pre-trained on the public COCO database can deal with target pedestrian areas detection well when the analyzed pedestrian area is large. When pedestrians are far away from camera, that is, the pedestrian area is small, the effect of target pedestrian areas detection with target pedestrian areas detection is usually unsatisfactory. In our proposed algorithm, a small target detection layer and Deformable-ELAN module are designed to improve the effect of small pedestrian areas detection. A total of 100 silhouette images and 100 dense optical flow images are selected to fine-tune the Gait-YOLO model. Experimental results in this paper show that the fine-tuned Gait-YOLO model with a small number of samples greatly improves the effect of pedestrian area extraction for both gait silhouette and dense optical flow images.

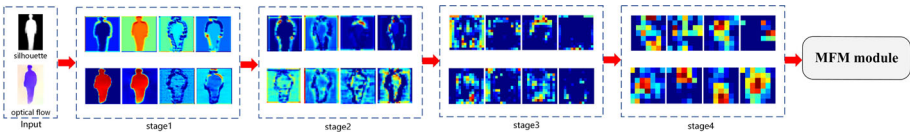

Fig. 7 Feature visualization extracted by CNNs on the gait images

3.4 Multi-modal fusion module

In this section, we propose a multi-modal fusion module (MFM) based on a squeeze-and-excitation (SE) scheme for discriminative multi-modal feature learning and classification. The aforementioned two kinds of gait representations, holistic silhouette and dense optical flow are used as the dual-channel input for feature fusion. The following of this section summarizes the main steps in recognizing an appearing frontal-view gait sequence using the multi-modal fusion module. The holistic flowchart of our proposed multi-modal fusion module is shown in Fig. 8.

First, in-depth gait characteristics underlying original holistic silhouettes and dense optical flow images are captured by using networks of feature extractors N_S and N_F, which have the same structure with different parameters. Both N_S and N_F contain four convolutional modules. Each convolutional module includes three 3×3 convolutional layers, one 1×1 convolutional layer, one pooling layer, and three SiLu activation function layers. The in-depth gait features underlying holistic silhouette and dense optical flow images are denoted as $S \in \mathbb{R}^{N_1 \times \ldots N_k \times C}$ and $F \in \mathbb{R}^{M_1 \times \ldots N_B \times C'}$, where N_k and M_k represent the spatial dimensions of the given gait representations, C and C' represents the number of channels for given gait representations. In order to show the process of feature learning with CNN models more intuitively, feature visualization is presented, in which the features obtained by the different stages of convolutional layers are visualized in Fig. 7 to demonstrate the learning capability of the low-, middle- and high-layer of the whole network.

The extracted holistic silhouettes maps S and dense optical flow maps F are further squeezed into a low-dimensional nonlinear space:

$$S_S(c) = \frac{1}{\prod_{i=1}^{K} N_i} \sum_{n_1,\ldots,n_K} S(n_1, \ldots, n_K, c) \tag{4}$$

$$S_F(c) = \frac{1}{\prod_{i=1}^{B} M_i} \sum_{m_1,\ldots,m_B} F(m_1, \ldots, m_B, c) \tag{5}$$

where S_S and S_F represent low-dimensional silhouette features and low-dimensional optical flow features after nonlinear mapping. The spatial information is then compressed into channel descriptors, and these two features are concatenated to obtain the concatenated low-dimensional features:

$$S_T = Concat(S_S, S_F) \tag{6}$$

Second, signal gating is applied with different calibration weights to two different modalities. The aforementioned low-dimensional features are then mapped to the global silhouette feature space E_S and the global optical flow feature space E_F using fully connected layers, respectively. An activation function is then applied to obtain output signal:

$$E_S = W_S S_S + b_S$$
$$E_F = W_F S_F + b_F \tag{7}$$

where W_S, b_S and W_F, b_F represent weights and biases corresponding to these two different modalities, respectively.

Third, the activated output signals E_S and E_F are used as weights for silhouette and optical flow feature maps in the channel dimension. The weighted feature maps are obtained by element-wise multiplication (Hadamard product) between the weights and the feature maps. The original feature maps and S and F are concatenated with the weighted feature maps after compression and activation through the SE attention module in the channel dimension. Then, the concatenated feature maps are compressed using a 1×1 convolution module to match the dimensions of original feature maps.

Fourth, the concatenated feature maps are encoded separately by the proposed gait feature encoder, called Gait-Encoder, to obtain the corresponding encoding vectors S'_E and F'_E. These two encoding vectors S'_E and F'_E are fused to obtain the final gait feature representation S'_T, which contains two different kinds of in-depth gait characteristics underlying silhouette and optical flow features.

Different from conventional transformer network, the proposed Gait-Encoder algorithm introduces residual structure for feature learning, which consists of three sub-modules: multi-head attention module, feed-forward network module, and additive module. Multi-head attention mechanism is used to capture long-range dependencies on the encoded input vectors: query (Q), key (K), and value (V), and learn the different internal structures of gait features:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{8}$$

$$MultiHead(Q, K, V) = Comcat(head_1, \ldots, head_h)W^0 \tag{9}$$

where $Q = W_q X$, $K = W_k X$, $V = W_V X$, and $head_i = Attention(Q, K, V)$.

The feedforward network module (FFNM) consists of two fully connected layers and a SiLu activation function layer. Linear transformation structure is the same for each layer, but different parameters are used, we have:

$$FFNM(x) = W_2\sigma(W_1 X) \tag{10}$$

where W_1 and W_2 represent parameter matrices of the two fully connected layers and $\sigma(\cdot)$ represents the sigmoid activation function.

Additive module is used and described as follows:

$$AM(x_m) = x_m + Attention(Q, K, V) \tag{11}$$

By using the FFNM, we can have

$$AM(x_f) = x_f + max(0, x_f W_1 + b_1)W_2 + b_2 \tag{12}$$

Different from existing other multi-modal learning algorithms in [32–34], the proposed MFM uses feature concatenation-based weighted fusion of contour features and optical flow

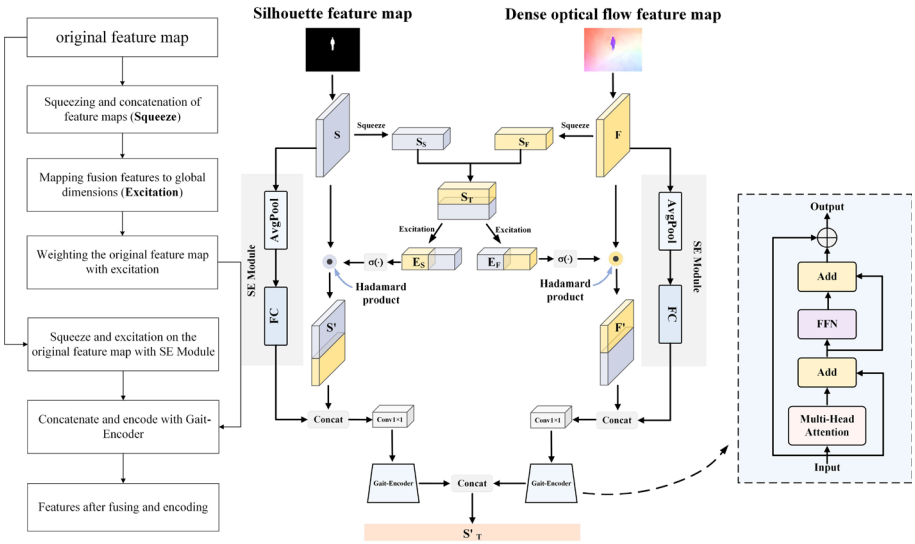

Fig. 8 Holistic flowchart of our proposed multi-modal fusion module

features. The two types of features are compressed and concatenated in space to obtain fused features. Compression units receive fused features and generate joint representations based on global information. Then, based on the weights and biases of each feature, the joint representations are mapped back to the original global feature dimension, achieving adaptive fusion. At the same time, the SE module is used to process the original features in the channel dimension, which are then merged with the fused features. This approach not only allows the model to focus more on valid information and suppress invalid features in both channel and spatial dimensions but also combines low-dimensional high-resolution information with high-dimensional high-semantic information to enable the model to effectively extract features without being affected by noise pollution.

The triplet loss is utilized in this paper as the loss function for the frontal-view gait recognition task, which can increase inter-class distance and decrease intra-class distance in the classification process:

$$L_{triplet} = max(D(F(i), F(k)) - D(F(i), F(j)) + m, 0) \tag{13}$$

where sample i and j belong to class A, sample k belongs to class B, $F(\cdot)$ represents the proposed gait feature extractors, $D(\cdot)$ represents Euclidean distance function, and m represents the margin of the triplet loss.

4 Experiments

In this section, two widely used gait databases, namely CASIA-B gait database [35] and OUMVLP gait database [36] are used in performance evaluation of our proposed method. Comparisons with existing other works are given in this section as well. Finally, ablation studies are conducted to determine contribution of each module of our proposed method.

Table 1 Averaged rank-1 recognition rates (%) on CASIA-B gait database under the normal walking condition

Training strategy	Gait features	Accuracy (%)
ST(24)	Holistic silhouette	67.7
	Optical flow	72.6
	Feature fusion	71.4
MT(62)	Holistic silhouette	78.1
	Optical flow	81.2
	Feature fusion	83.3
LT(74)	Holistic silhouette	87.8
	Optical flow	84.5
	Feature fusion	90.6

In the training phase, the batch all (BA) algorithm is used for parameter training with the batch size set to (p, k), where p represents the number of subjects, and k represents the number of samples for each subject. Our proposed method adopts a dual-channel input of silhouette images and dense optical flow maps. The input sizes of both silhouette and dense optical flow maps are set to 64×44. The optimizer used is Adam with a learning rate of 0.005. The margin for the triplet loss is set to 0.2, and the batch size is set to (8, 8) using the BA sampling strategy. The experimental results in this section are reported in term of rank-1 accuracy. All algorithms and experiments are implemented by using Python platform and a server with NVIDIA GeForce 3080 GPU.

4.1 Experiments on CASIA-B Gait Database

CASIA-B database contains walking videos and silhouette images of 124 subjects, with 11 different view angles (0°-180°) for each subject, and 10 sequences for each view. Each subject has 3 different walking states, including normal (NM#01-06), carrying a backpack (BG#01-02), and wearing a jacket or coat (CL#01-02). Three training strategies, namely small-sample training (ST), medium-sample training (MT) and large-sample training (LT), are used in our experiments. Under these three training strategies, the first 24, 62, and 74 subjects are assigned into the training set, respectively.

4.1.1 Experimental Results Under the Normal Walking Condition

Two types of experiments on CASIA-B database are carried out to evaluate the performance of the proposed gait recognition algorithms. The first type is gait recognition without any walking variations, that is, both gallery and probe sequences are under the normal walking condition. In our experiments, the first three sequences (NM#01-03) are regarded as gallery, and the remaining three sequences (NM#04-06) as probes. Detailed experimental results are presented in Table 1.

It can be observed from the experimental results that (1) the feasibility of holistic silhouette and dense optical flow as identifiers of individuals is confirmed; (2) the power of discriminability provided by holistic silhouette is close to that provided by dense optical flow; (3) recognition rates using the proposed multi-modal fusion algorithm are better than that using single modality. The combined use of multi-modal walking characteristics can surely

Table 2 Averaged rank-1 recognition rates (%) on CASIA-B gait database under different walking conditions

Training strategy	NM	BG	CL	Avg
ST(24)	75.4	64.9	45.6	62.0
MT(62)	91.1	84.5	67.4	81.0
LT(74)	96.9	93.5	77.8	89.4

involve more informative features, which can improve the generalization performance of the model.

4.1.2 Experimental Results Under Changes of Walking Conditions

The second type of experiment is gait recognition under changes of walking conditions. The first four sequences (NM#01-04) are regarded as gallery, and the remaining sequences (NM#05-06, BG#01-02, CL#01-02) as probes. Detailed experimental results are presented in Table 2. We can observe from the experimental results that our proposed algorithm can still obtain promising results even under changes of different walking conditions. The extracted in-depth gait characteristics underlying holistic silhouette and dense optical flow are not sensitive to the variations of carrying and clothing statuses. Moreover, the proposed method fuses two different kinds of gait features, which provides more comprehensive information to develop a robust recognition system against variations of carrying and clothing statuses.

In order to further evaluate the effectiveness of MFM module, experiments under different feature fusion schemes are conducted. The first method uses weighted summation of the two feature map matrices, with a weight ratio of 1:1 (Add Only), the second method directly concatenates the silhouettes and dense optical flow maps (concatenate only), The third method uses the proposed MFM module for feature map fusion. Figure 9 shows the convergence of training losses under different fusion methods. Faster convergence and lower loss values indicate better performance. It can be observed that the proposed MFM module consistently obtains the best convergence results under three different walking conditions NM, BG, and CL. Particularly, under the BG and CL conditions, the training performance using the MFM module far surpassed that of simple concatenation and addition. This suggests that the MFM module exhibits superior capability in extracting gait motion information, particularly when confronting with significant changes in walking conditions. Moreover, the introduction of the SE (squeeze-and-excitation) attention module, fully connected layers, and an encoder within the MFM (multimodal fusion module) does lead to a slight increase in the number of parameters. However, due to its single-layer structure and the parameter-friendly nature of the SE attention mechanism in shallow networks, along with the improved convergence speed and detection accuracy brought about by the introduction of the MFM structure, the cost-effectiveness of the MFM structure far surpasses that of simple concatenation or addition-based methods.

4.1.3 Ablation Study

In this section, ablation study is conducted to evaluate the improvement of the proposed three main modules, including Gait-YOLO pedestrian detection method, multimodel feature fusion module (MFM), and gait-encoder (GE) algorithm, on frontal-view gait recognition performance. Detailed experimental results are shown in Table 3, which reports the averaged rank-1 accuracy with the training strategy of LT.

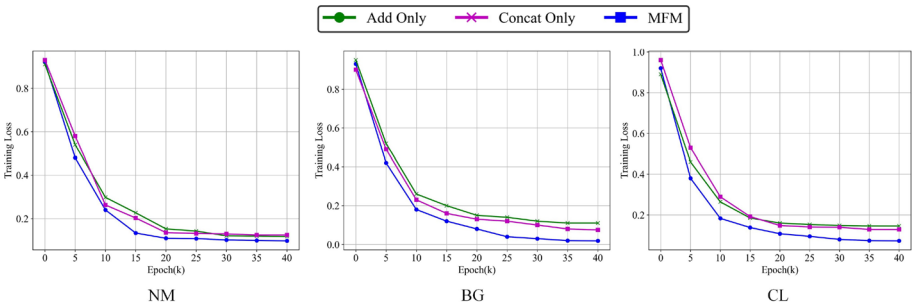

Fig. 9 Training loss curves with different fusion strategies

Table 3 Ablation studies on CASIA-B gait database

Experiment	Gait-YOLO	MFM	GE	NM	BG	CL
A				86.3	78.0	55.7
B	✓			93.5	81.6	62.1
C	✓	✓		92.2	82.7	65.1
D	✓	✓	✓	96.9	93.5	77.8

This paper proposes an improved YOLOv7-based pedestrian region detection method, which includes a small object detection layer and a Deformable-ELAN module with a global receptive field to efficiently detect frontal-view pedestrian regions. By comparing the results of experiments A and B (with the Gait-YOLO method) in Table 3, significant improvement can be observed, indicating that the improved YOLOv7 algorithm can effectively extract pedestrian regions, exclude background noise and facilitate accurate gait recognition.

Experiments B and C are carried out to investigate the impact of the multi-modal fusion module. Compared to experiment B, experiment C shows a 0.7% decrease in recognition accuracy only under NM condition, while increase in accuracy of 1.1% and 3.0% are observed under BG and CL conditions, respectively. This can be attributed to the fact that multi-modal fusion module is aimed at enhancing the diversity of semantic information from different aspects of gait characteristics. The model is suitable to handle variations in pedestrian appearance due to carrying backpacks or wearing coats, thereby enabling accurate identity discrimination.

Gait feature encoder is proposed to encode the fused gait feature vector, and its effectiveness is verified by comparing the results before (experiment C) and after (experiment D) adding the Gait-encoder (GE), which shows improvements in accuracy rates by 3.9%, 1.8%, and 2.3% for these three walking conditions, respectively. The proposed gait feature encoder uses a multi-head attention module to extract multi-motion features, not only in the temporal dimension for short-, medium-, and long-term features, but also for modeling the rich information between different input features and simulating multi-motion situations without depth superposition. Its powerful global modeling capability enables the gait feature encoder to effectively mine the relationship between different blocks of pixel regions, and fully characterize the gait semantic information.

Table 4 Comparison with state-of-the-art frontal-view gait recognition methods on CASIA-B gait database

Method	NM	BG	CL	Avg
CNN-LB [41]	82.6	64.2	37.7	61.5
GaitSet [17]	90.8	83.8	61.4	78.7
GaitPart [38]	94.1	89.1	70.7	84.6
GLN [39]	93.2	91.1	70.6	85.0
MT3D [40]	95.7	91.0	76.0	87.6
Proposed method	96.9	93.5	77.8	89.4

4.1.4 Computational Complexity

For the image preprocessing step, this paper employs Gait-YOLO algorithm for feature extraction from gait pedestrian regions. The Gait-YOLO model has 48.28 million parameters and a computational complexity of 106.2 GFLOPs, slightly exceeding the original YOLOv7, which has 36.49 million parameters and 103.5 GFLOPs. This is because the addition of a small-object detection layer in the YOLOv7 neck region enhance recognition accuracy, resulting in a minor increase in both parameter count and computational complexity. Furthermore, this paper utilizes a custom-designed four-stage convolutional neural network as the backbone for extracting gait features from silhouette and dense optical flow sequences. This backbone network have shallower depth, but only contains 1.2 million parameters, significantly fewer than mainstream deep neural networks like ResNet and VGG. The parameter count of the MFM (multimodal fusion module) can be considered negligible in comparison. Training is conducted on an Nvidia 3080 graphics processing unit (GPU). For one training iteration of a motion pattern, the average time is 10 s, while the average time for a single inference prediction is 0.7 s.

4.1.5 Comparisons with Other Existing Frontal-View Methods

This section further investigates performance comparisons with other existing frontal-view gait recognition methods on CASIA-B database. Several mainstream gait recognition methods are selected for comparison: CNN-LB [37], GaitSet [17], GaitPart [38], GLN [39], and MT3D [40]. Detailed experimental results of comparison can be seen in Table 4.

From the experimental results, we can observe that: (1) Our proposed method outperforms existing frontal-view gait recognition methods. For the case of NM condition, our proposed method outperforms the best-performing MT3D method by 1.2%. For the case of BG condition, our proposed method outperforms GLN by 2.4%. For the case of CL condition, our proposed method outperforms MT3D by 1.8%. The reported significant improvements of our method under the BG and CL walking conditions are attributed to the introduction of dense optical flow features, which increase the amount of information captured for gait representations. Additionally, the motion characteristics of different body parts represented by dense optical flow are robust to changes in clothing and carrying, resulting in good robustness performance under BG and CL conditions.

4.1.6 Comparisons with Other Existing Single-Modal Methods

This section mainly investigates performance comparisons with other existing single-modal methods. Detailed comparison results on the CASIA-B database are shown in Table 5. From

Table 5 Comparison with state-of-the-art single-modal gait recognition methods on CASIA-B gait database

Method	NM	BG	CL	Avg
LF + AVG [42]	71.4	13.1	20.2	34.9
LF + DTW [42]	61.9	21.4	25.0	36.1
LF + oHMM [42]	63.8	19.7	22.6	35.3
GEI + PCA + LDA [43]	91.2	44.5	23.7	53.1
GPPE [44]	93.4	56.1	22.4	57.3
GEnI [45]	92.3	56.4	27.2	58.6
Proposed method	96.9	93.5	77.8	89.4

Table 6 Comparison with state-of-the-art multi-modal gait recognition methods on CASIA-B gait database

Method	NM	BG	CL	Avg
MissGait [46]	99.6	89.5	45.5	78.2
mmGaitSet [47]	95.6	91.4	77.6	88.2
GaitPartHybrid [48]	96.4	89.1	75.4	87.0
Proposed method	96.9	93.5	77.8	89.4

the experimental results, we can see that the recognition performance of our proposed multi-modal method is superior to other existing single-modal methods, especially for the case of walking condition variations. Our proposed method fuses two different modals for frontal-view gait sequences, leading to a higher recognition accuracy compared with single-modal gait recognition methods. The multi-modal information fusion can provide an comprehensive characterization of human walking, avoiding the great drops of recognition rates when walking conditions change.

4.1.7 Comparisons with Other Existing Multi-modal Methods

Comparisons with other existing multi-modal gait recognition methods are given in this section as well on the CASIA-B gait database. Table 6 compares the average recognition results of the proposed method under the aforementioned three walking conditions with the ones of existing other gait recognition that used different combinations of modalities, including MissGait [46], mmGaitSet [47], and GaitPartHybrid [48]. The modalities used in these methods include grayscale images, optical flow images, silhouette images, pose heatmaps, and skeleton images.

It can be observed from the experimental results that, compared with MissGait method using grayscale images and optical flow, our method is slightly lower than MissGait by 2.7% under the NM walking condition, but outperforms MissGait by 4.0% and 32.3% under the BG and CL conditions, respectively. This is because the silhouette images carry less noise than grayscale images, thus enhancing the robustness of our method to clothing and carrying variations. Compared with GaitPartHybrid, our method achieves 0.5%, 4.4%, and 2.4% higher accuracy in these three walking conditions, respectively. This result indicates that dense optical flow has a more powerful discriminability than skeleton images under clothing and carrying variations. When the appearance of pedestrians changes, the motion trends represented by dense optical flow are often more stable.

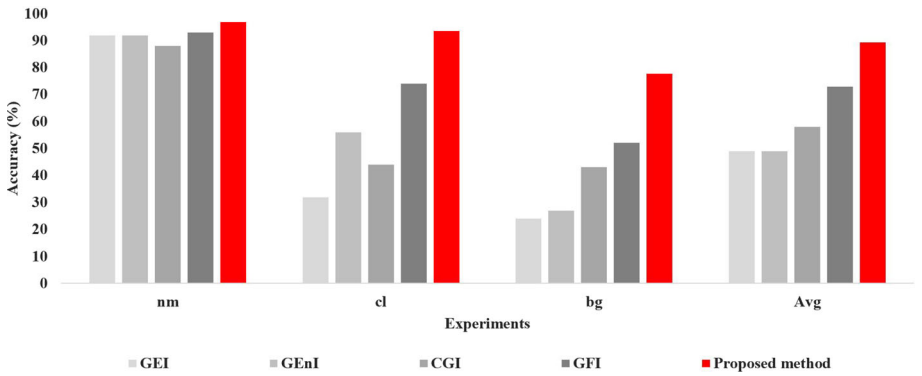

Fig. 10 Comparison with state-of-the-art gait representation methods on CASIA-B gait database

4.1.8 Comparisons with Other Existing Gait Representation Methods

We further compare our proposed method with other existing gait representation methods, namely GEI [49], GEnI [50], CGI [51], GFI [10], on the CASIA-B gait database. It can be observed form the experimental results in Fig. 10 that our proposed frontal-view gait representation is not inferior to other widely used gait representations for the NM condition. Moreover, our proposed method makes full use of holistic silhouette and dense optical flow to enhance the robustness against a wide range of walking condition variations, avoiding great drops of recognition rates due to the variations of walking conditions.

4.1.9 Comparisons with Other Existing Deep-Learning-Based Methods

Comparisons with other existing deep-learning-based methods are investigated in this section on the CASIA-B database. Algorithms in [41], [52], [53] and [54] are implemented, and detailed experimental results can be found in Fig. 11. It can be seen that the proposed method outperforms other existing deep-learning-based methods in terms of average recognition rates. Although algorithms in Refs. [41] and [52] can achieve higher recognition rates under the NM conditions, they suffer from great drops of recognition rates when walking condition change. In this sense, our proposed method makes full use of holistic silhouette and dense optical flow, which is expected to provide reliable performance against different kinds of walking conditions.

4.2 Experiments on OUMVLP Gait Database

This section further evaluates performance of our proposed method on the OUMVLP gait database. The OUMVLP gait database is known as the largest public gait database, containing 10,307 individuals, with 14 views for each subject. Each subject has 2 different walking sequences, denoted as 00 and 01 sequence. In our experiments, a total of 5153 subjects are assigned into training set and the remaining subjects for testing. In the testing phase, the 01 sequence is used as the gallery, and the 00 sequence is used as the probe. The training and testing strategies are the same as those of Gaitset and GaitPart. We compare our proposed method with the GEINet, GaitSet, GaitPart, GLN method, and detailed experimental results are shown in Fig. 12. The experimental results have demonstrated the feasibility of the pro-

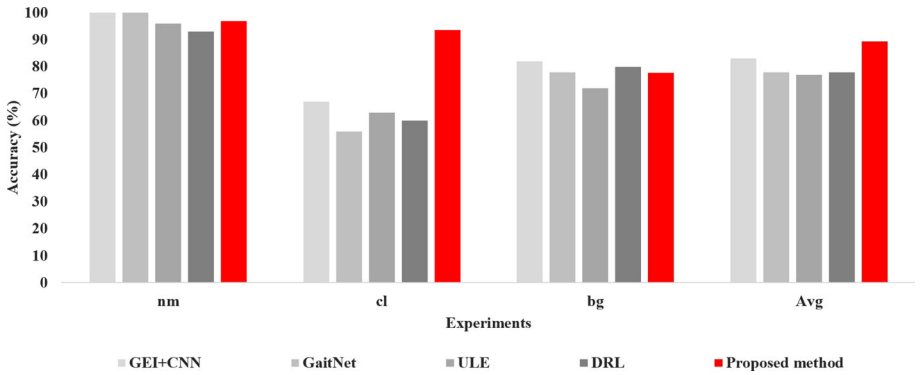

Fig. 11 Comparison with state-of-the-art deep-learning-based gait recognition methods on CASIA-B gait database

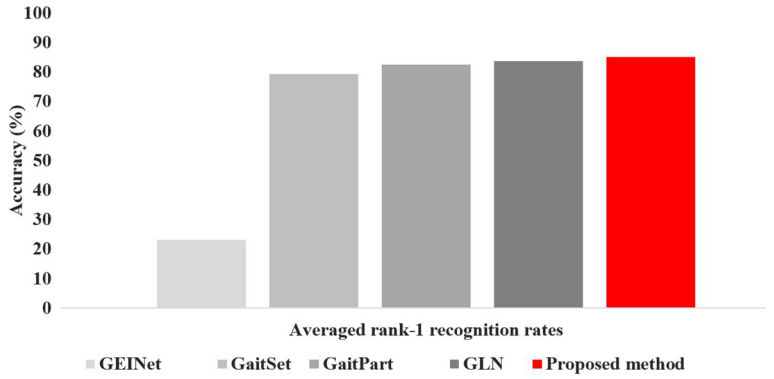

Fig. 12 Comparison with state-of-the-art gait representation methods on OUMVLP gait database

posed method in the frontal-view gait recognition. Our proposed method can work well in recognizing gait patterns from a frontal view.

4.3 Experiments on CASIA-A Gait Database

Recognition performance of the proposed method is finally evaluated on CASIA-A gait database [55], in which 20 different subjects and 40 gait sequences under normal walking condition are involved. Compared with CASIA-B database, CASIA-A gait database is collected in outdoor environments, and is suitable to investigate generalization performance of our proposed method. A total of 40 frontal-view gait sequences are used in our experiments. The first sequence is assigned into the probe set and the remaining sequence into gallery set. We transfer the knowledge of deep learning model trained by the CASIA-B database to evaluate the generalization performance on CASIA-A gait database. The rank-1 recognition rate is 95.0%, indicating that the proposed method can achieves excellent performance in cross-domain gait recognition.

5 Discussion

From the aforementioned experimental results, the following observations can be obtained:

- Our proposed method facilitates the applications of gait-based individual identification in practice. Superior performance can be obtained compared with other existing gait recognition methods. It is expected to provide a more practical tool in real-world environment for individual identification, even in secure narrow corridor.
- Our proposed method still achieves reliable performance compared with existing other methods when the testing walking conditions are different from the corresponding training conditions. Despite that existing other methods can obtain similar recognition rates [56, 57], the proposed method can still work well against different walking condition variations. The fusion of silhouette and dense optical flow features from the frontal-view gait sequences provides a comprehensive representation of gait dynamics, making it insensitive to variations in clothing, carrying status, walking speed, and temporal changes.
- Our proposed method enhances the recognition accuracy of single modality. The proposed silhouette and dense optical flow features are designed to complement each other in the recognition task. Silhouette images can capture the missing contour information present in dense optical flow maps, while dense optical flow maps provide insights into the internal motion patterns absent in silhouette images. Multi-modal feature fusion can be achieved by using a multi-modal fusion module (MFM) that integrates features from different aspects, yielding superior results compared to single modality. Despite that much progress has been reported on multi-modal feature learning [58–60], the proposed MFM module exhibits superior capability in extracting gait motion information, particularly when confronting with significant changes in walking conditions.
- The proposed object region extraction algorithm, namely Gait-YOLO, partially addresses the challenge of varying target receptive fields. Moreover, a viable strategy is introduced for small target detection, involving the incorporation of small target detection layers at the channel level and the introduction of DCNv3 convolutional layers at the structural level. These methods simultaneously enlarge the receptive fields and enhance long-distance modeling capabilities, significantly improving the recognition accuracy.
- In previous researches, the incorporation of silhouette and dense optical flow feature sequences into end-to-end networks posed challenges on complex feature extraction steps. Our proposed method introduces the Gait-YOLO scheme to automatically extract pedestrian regions from raw walking video sequences, simplifying the feature preprocessing steps and enhancing the accuracy of recognition and regression. Additionally, convolutional neural networks are employed to further extract features underlying both silhouettes and dense optical flow, followed by the utilization of the designed multi-modal fusion module (MFM) to fuse these two types of features, resulting in a more robust representation of gait patterns.

6 Conclusion

This paper presents a novel framework for frontal view gait recognition, which investigates multi-modal feature representations and learning. Compared with existing other gait recognition methods, the strength of this paper is that, we propose a discriminative feature representation scheme by fusing holistic silhouette and dense optical flow from frontal-view walking sequences. Another strength of our proposed method lies in that we introduce a

multi-modal feature fusion module to explore the intrinsic connections between silhouette and dense optical flow features by using squeeze and excitation operations at the channel and spatial levels. Gait feature encoder is further used to extract global walking characteristics, enabling efficient multi-modal information fusion. Experimental results on the CASIA-B and OUMVLP database demonstrate that promising recognition results can be obtained. Future work will focus on the improvement of algorithm robustness against a wider range of walking conditions.

Acknowledgements This research work is supported by the National Natural Science Foundation of China (Nos. T2341019, 61803133), and by the Guangzhou Basic and Applied Basic Research Project (Nos. 202201010346, 2023A04J0347), and by the Guangdong Basic and Applied Basic Research Foundation (Nos. 2021A1515012635, 2023A1515011245), and by the Open Foundation of the Guangdong Provincial Key Laboratory of Electronic Information Products Reliability Technology.

Author Contributions Muqing Deng: algorithm design; Zebang Zhong: software and validation; Yi Zou: formal analysis and visualization; Yanjiao Wang: conceptualization and supervision; Kaiwei Wang: data curation; Junrong Liao: review and editing.

Declarations

Statements and Declarations We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work, there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

Data Availability Statement Data openly available in a public repository. Data used in the present study are publicly available, and ethical approval and informed consent can be obtained in each original study.

References

1. Deng M, Fan T, Cao J, Fung S-Y, Zhang J (2020) Human gait recognition based on deterministic learning and knowledge fusion through multiple walking views. J Frankl Inst 357(4):2471–2491
2. Zhao A, Dong J, Li J, Qi L, Zhou H (2021) Associated spatio-temporal capsule network for gait recognition. IEEE Trans Multimedia 24:846–860
3. Zhang Z, Tran L, Liu F, Liu X (2022) On learning disentangled representations for gait recognition. IEEE Trans Pattern Anal Mach Intell 44(1):345–360
4. Ryu J, Kamata S (2011) Front view gait recognition using spherical space model with human point clouds. In: 2011 18th IEEE international conference on image processing. IEEE, pp 3209–3212
5. Soriano M, Araullo A, Saloma C (2004) Curve spreads—a biometric from front-view gait video. Pattern Recognit Lett 25(14):1595–1602
6. Tahmoush D, Silvious J (2009) Radar micro-Doppler for long range front-view gait recognition. In: 2009 IEEE 3rd international conference on biometrics: theory, applications, and systems. IEEE, pp 1–6
7. Barnich O, Van Droogenbroeck M (2009) Frontal-view gait recognition by intra-and inter-frame rectangle size distribution. Pattern Recognit Lett 30(10):893–901
8. Matovski DS, Nixon MS, Mahmoodi S, Carter JN (2011) The effect of time on gait recognition performance. IEEE Trans Inf Forensics Secur 7(2):543–552

9. Bouchrika I, Nixon MS (2008) Exploratory factor analysis of gait recognition. In: 2008 8th IEEE international conference on automatic face & gesture recognition. IEEE, pp. 1–6
10. Lam TH, Cheung KH, Liu JN (2011) Gait flow image: a silhouette-based gait representation for human identification. Pattern Recognit 44(4):973–987
11. Liao R, Cao C, Garcia EB, Yu S, Huang Y (2017) Pose-based temporal-spatial network (PTSN) for gait recognition with carrying and clothing variations. In: Biometric recognition: 12th Chinese conference, CCBR 2017, Shenzhen, China, October 28–29, 2017, proceedings 12. Springer, pp 474–483
12. Zhang C, Liu W, Ma H, Fu H (2016) Siamese neural network based gait recognition for human identification. In: 2016 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 2832–2836
13. Bari AH, Gavrilova ML (2019) Artificial neural network based gait recognition using Kinect sensor. IEEE Access 7:162708–162722
14. Battistone F, Petrosino A (2019) TGLSTM: a time based graph deep learning approach to gait recognition. Pattern Recognit Lett 126:132–138
15. Teepe T, Gilg J, Herzog F, Hörmann S, Rigoll G (2022) Towards a deeper understanding of skeleton-based gait recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1569–1577
16. Zhao A, Li J, Ahmed M (2020) SpiderNet: a spiderweb graph neural network for multi-view gait recognition. Knowl-Based Syst 206:106273
17. Chao H, He Y, Zhang J, Feng J (2019) GaitSet: regarding gait as a set for cross-view gait recognition. Proc. AAAI Conf. Artif. Intell. 33:8126–8133
18. Shiraga K, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2016) GeiNet: view-invariant gait recognition using a convolutional neural network. In: 2016 international conference on biometrics (ICB). IEEE, pp 1–8
19. Tong S, Fu Y, Ling H (2019) Cross-view gait recognition based on a restrictive triplet network. Pattern Recognit Lett 125:212–219
20. Vaezi Joze HR, Shaban A, Iuzzolino ML, Koishida K (2020) MMTM: multimodal transfer module for CNN fusion. In: 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 13286–13296. https://doi.org/10.1109/CVPR42600.2020.01330
21. Papavasileiou I, Qiao Z, Zhang C, Zhang W, Bi J, Han S (2021) GaitCode: gait-based continuous authentication using multimodal learning and wearable sensors. Smart Health 19:100162
22. Marín-Jiménez MJ, Castro FM, Delgado-Escaño R, Kalogeiton V, Guil N (2021) UGaitNet: multimodal gait recognition with missing input modalities. IEEE Trans Inf Forensics Secur 16:5452–5462
23. Liu X, Chen M, Liang T, Lou C, Wang H, Liu X (2022) A lightweight double-channel depthwise separable convolutional neural network for multimodal fusion gait recognition. Math Biosci Eng 19:1195–1212
24. Li G, Guo L, Zhang R, Qian J, Gao S (2023) TransGait: multimodal-based gait recognition with set transformer. Appl Intell 53(2):1535–1547
25. Goffredo M, Carter JN, Nixon MS (2008) Front-view gait recognition. In: 2008 IEEE second international conference on biometrics: theory, applications and systems . IEEE, pp 1–6
26. Chattopadhyay P, Roy A, Sural S, Mukhopadhyay J (2014) Pose depth volume extraction from RGB-D streams for frontal gait recognition. J Vis Commun Image Represent 25(1):53–63
27. Zulcaffle TMA, Kurugollu F, Crookes D, Bouridane A, Farid M (2019) Frontal view gait recognition with fusion of depth features from a time of flight camera. IEEE Trans Inf Forensics Secur 14(4):1067–1082
28. Rashmi M, Guddeti RMR (2022) Human identification system using 3D skeleton-based gait features and LSTM model. J Vis Commun Image Represent 82:103416
29. Chattopadhyay P, Sural S, Mukherjee J (2013) Gait recognition from front and back view sequences captured using kinect. In: International conference on pattern recognition and machine intelligence. Springer, pp 196–203
30. Liao R, Yu S, An W, Huang Y (2020) A model-based gait recognition method with body pose and human prior knowledge. Pattern Recognit 98:107069
31. Mao M, Song Y (2020) Gait recognition based on 3D skeleton data and graph convolutional network. In: 2020 IEEE international joint conference on biometrics (IJCB). IEEE, pp 1–8
32. Zhang J, Yang J, Yu J, Fan J (2022) Semisupervised image classification by mutual learning of multiple self-supervised models. Int J Intell Syst 37(5):3117–3141
33. Yu J, Tan M, Zhang H, Rui Y, Tao D (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell 44(2):563–578
34. Zhang J, Cao Y, Wu Q (2021) Vector of locally and adaptively aggregated descriptors for image feature representation. Pattern Recognit 116:107952

35. Yu S, Tan D, Tan T (2006) A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th international conference on pattern recognition (ICPR'06), vol 4. IEEE, pp 441–444
36. Takemura N, Makihara Y, Muramatsu D, Echigo T, Yagi Y (2018) Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans Comput Vis Appl 10(1):1–14
37. Wu Z, Huang Y, Wang L, Wang X, Tan T (2016) A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans Pattern Anal Mach Intell 39(2):209–226
38. Fan C, Peng Y, Cao C, Liu X, Hou S, Chi J, Huang Y, Li Q, He Z (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
39. Hou S, Cao C, Liu X, Huang Y(2020) Gait lateral network: learning discriminative and compact representations for gait recognition. In: Computer vision—ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part IX. Springer, pp 382–398
40. Lin B, Zhang S, Bao F (2020) Gait recognition with multiple-temporal-scale 3D convolutional neural network. In: Proceedings of the 28th ACM international conference on multimedia, pp 3054–3062
41. Wu Z, Huang Y, Wang L, Wang X, Tan T (2016) A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans Pattern Anal Mach Intell 39(2):209–226
42. Hu M, Wang Y, Zhang Z, Zhang D, Little JJ (2013) Incremental learning for video-based gait recognition with LBP flow. IEEE Trans Cybern 43:77–89
43. Sarkar S, Phillips PJ, Liu Z, Vega IR, Grother P, Bowyer KW (2005) The human ID gait challenge problem: data sets, performance, and analysis. IEEE Trans Pattern Anal Mach Intell 27:162–177
44. Jeevan M, Jain N, Hanmandlu M, Chetty G (2013) Gait recognition based on gait pal and pal entropy image. In: IEEE international conference on image processing, pp 4195– 4199
45. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image. In: International conference on imaging for crime detection and prevention, pp 1–6
46. Delgado-Escano R, Castro FM, Guil N, Kalogeiton V, Marin-Jimenez MJ (2021) Multimodal gait recognition under missing modalities. In: 2021 IEEE international conference on image processing (ICIP). IEEE, pp 3003–3007
47. Zhao L, Guo L, Zhang R, Xie X, Ye X (2022) mmGaitSet: multimodal based gait recognition for countering carrying and clothing changes. Appl Intell 52(2):2023–2036
48. Cai N, Feng S, Gui Q, Zhao L, Pan H, Yin J, Lin B (2021) Hybrid silhouette-skeleton body representation for gait recognition. In: 2021 13th international conference on intelligent human-machine systems and cybernetics (IHMSC). IEEE, pp 216–220
49. Man J, Bhanu B (2006) Individual recognition using gait energy image. IEEE Trans Pattern Anal Mach Intell 28(2):316–322
50. Bashir K, Xiang T, Gong S (2009) Gait recognition using gait entropy image. In: International conference on imaging for crime detection and prevention, pp 1– 6
51. Wang C, Zhang J, Wang L, Pu J, Yuan X (2012) Human identification using temporal information preserving gait template. IEEE Trans Pattern Anal Mach Intell 34:2164–2176
52. Song C, Huang Y, Huang Y, Jia N, Wang L (2019) GaitNet: an end-to-end network for gait based human identification. Pattern Recognit 96:106988
53. Zhang S, Wang Y, Li A (2021) Cross-view gait recognition with deep universal linear embeddings. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9095–9104
54. Yao L, Kusakunniran W, Zhang P, Wu Q, Zhang J (2022) Improving disentangled representation learning for gait recognition using group supervision. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2022.3171961
55. Wang L, Tan T, Hu W, Ning H (2003) Automatic gait recognition based on statistical shape analysis. IEEE Trans Image Process 12(9):1120–1131
56. Zhang J, Cao Y, Wu Q (2021) Vector of locally and adaptively aggregated descriptors for image feature representation. Pattern Recognit 116:107952
57. Zhang J, Yang J, Yu J, Fan J (2022) Semisupervised image classification by mutual learning of multiple self-supervised models. Int J Intell Syst 37(5):3117–3141
58. Hong C, Yu J, Zhang J, Jin X, Lee K-H (2018) Multimodal face-pose estimation with multitask manifold deep learning. IEEE Trans Ind Inf 15(7):3952–3961
59. Yu J, Tan M, Zhang H, Rui Y, Tao D (2019) Hierarchical deep click feature prediction for fine-grained image recognition. IEEE Trans Pattern Anal Mach Intell 44(2):563–578
60. Liu J, Zhang L, Zhu S, Liu B, Liang Z, Yang S (2022) Exploring complex dependencies for multi-modal semantic trajectory prediction. Neural Process Lett 54:961–985