IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 October 2023, accepted 5 November 2023, date of publication 8 November 2023,
date of current version 17 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3331395

RESEARCH ARTICLE

Cross-View Gait Recognition Model Combining Multi-Scale Feature Residual Structure and Self-Attention Mechanism

JINGXUE WANG[1,2], (Member, IEEE), JUN GUO[1], (Member, IEEE), AND ZHENGHUI XU[1], (Member, IEEE)

[1]School of Geomatics, Liaoning Technical University, Fuxin 123000, China
[2]Collaborative Innovation Institute of Geospatial Information Service, Liaoning Technical University, Fuxin 123000, China

Corresponding author: Jun Guo (472120801@stu.lntu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Project 41871379, in part by the Liaoning Revitalization Talents Program under Project XLYC2007026, and in part by the Fundamental Applied Research Foundation of Liaoning Province under Grant 2022JH2/101300273.

This work involved human subjects or animals in its research. The authors confirm that all human/animal subject research procedures and protocols are exempt from review board approval.

ABSTRACT In the cross-view condition, the gait recognition rate caused by the vastly different gait silhouette maps is substantially reduced. To improve the accuracy of gait recognition under cross-view conditions, this paper proposes a cross-view gait recognition network model combining multi-scale feature residual module (MFRM) and self-attention (SA) mechanism based on Generative Adversarial Network (GAN). First, the local and global feature information in the input gait energy image is fully extracted using the MFRM. Then, the SA mechanism module is used to adjust the information of channel dimensions and capture the association between feature information and is introduced into both the generator and discriminator. Next, the model is trained using a two-channel network training strategy to avoid the pattern collapse problem during training. Finally, the generator and discriminator are optimized to improve the quality of the generated gait images. This paper conducts experiments using the CASIA-B and OU-MVLP public datasets. The experiments demonstrate that the MFRM can better obtain the local and global feature information of the images. The SA mechanism module can effectively establish global dependencies between features, so that the generated gait images have clearer and richer detail information. The average Rank-1 recognition accuracies of the results in this paper reach 91.1% and 97.8% on the two datasets respectively, which are both better than the current commonly used algorithms, indicating that the network model in this paper can well improve the gait recognition accuracy across perspectives.

INDEX TERMS Cross-view, gait recognition, residual module, self-attention mechanism, two-channel networks.

I. INTRODUCTION

With the rapid development of the information age worldwide, the modernization ability of national governance can be strengthened by identifying personal identity rapidly and accurately through biometric identification technology. Gait recognition, as a novel biometric identification technology in recent years, is capable of personal identification by their walking posture. Compared with biometric identification technologies such as the human face, fingerprint, and iris, gait recognition has the advantages of long distance [1], no contact [2], no need for cooperation [3], and camouflage prevention. Therefore, gait recognition technology has extensive application prospects and economic values in such fields as security monitoring [4], criminal investigation [5], human-computer interaction, and medical diagnosis [6]. In actual

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

scenes, however, the accuracy of gait recognition is affected by many factors like clothing changes [7], walking speed [8], carrying condition [9], and cross-viewing angle [10]. Clothing changes, walking speed, and belongings are all subjective factors, which cannot easily be optimized. The cross-view problem belongs to an objective factor of instrument shooting, which can be well optimized through models to improve the accuracy of gait recognition. Therefore, the cross-view angle is regarded as the factor affecting gait recognition most [11], [12], and it is also a research hotspot and mainstream issue in the field of gait recognition.

At present, the research on cross-view gait recognition through traditional methods can be roughly divided into two categories: one is based on model matching and the other is based on appearance matching. The model matching-based method [13], [14] performs modeling on the basis of the skeletal structure of the human body in the high-resolution gait image and reaches the goal of personal identification by extracting the traveling trajectories of the subject. Despite a good gait identification effect under cross-view scenes, this method needs to calculate a complex model, occupies large computing resources, and proposes high requirements for the quality of gait images. Comparatively, the method based on appearance matching only takes the gait silhouette map obtained from the low-resolution gait image as the input and judges the identity of the observed object by identifying the distinguishing feature information. Generally, this method takes a periodic gait sequence set [15] or energy-like graphs [16], [17], [18], [19] as the input. Gait energy image (GEI) in the energy-like image is a normalized image of the observed object in a walking period, which can effectively represent the relative location change of body silhouette in a period, so it has been widely used. Taking the gait silhouette of the observed object as the input, the method based on appearance matching is more sensitive to the change in individual appearance, especially when walking with a backpack or wearing a coat, which is more difficult to identify.

In recent years, the rapid development of deep learning technology provides a new idea for cross-view gait recognition. At present, there are many deep learning network models based on traditional appearance-matching technology. These network models can be roughly divided into view transformation model [20], [21], feature fusion model [22], [23], and deep neural network models [24], [25]. The view transformation model can realize personal identification using a singular value decomposition feature matrix to realize identity recognition, but it cannot use the feature information from all views at the same time, and a large number of training samples are usually needed for modeling. Comparatively, the feature fusion model can make full use of different feature information for joint modeling, thereby achieving the purpose of recognition. Among them, many researchers pay more attention to the method which can fuse and aggregate the spatial-temporal information from the skeletons and silhouettes [26], [27]. These two types of features are complementary, so the performance of these models becomes better after feature fusion. However, fusing the two kind of features is not simple, it requires complex modeling, high-precision feature alignment, and stronger computing power [28]. In addition, there are large semantic gaps between different types of feature information [29], and these semantic gaps cannot be completely eliminated after feature fusion. Therefore, there is often considerable noise in the fused model, greatly influencing the recognition performance of the model. Different from the first two models, the deep neural network model is based on commonly used network architectures, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and generative adversarial networks (GANs). The deep neural network model extracts, fuses, and classifies important feature information through different modules and finally identifies the identity through comparisons. Given that it can be integrated with different modules, the deep neural network model can effectively extract the deep and shallow features of the research object, make full use of the feature information from multiple views and scales, and minimize the impact of the semantic gap.

Deep neural network models often incorporate the ideas of feature fusion and view transformation and elevate the recognition rate by improving the model structure or adding modules. Hu et al. [30] integrated a multi-branch residual structure into the CNNs, which facilitated the model to make full use of the more representative feature information of the input image. Zhai [31] used an autoencoder (AE) to separate identity features and view features from GEI for coding recombination and reconstructed the GEI for identity recognition. Wang et al. [32] constrained the generator using multiple loss functions on the basis of GANs to improve the performance of the model. Following the idea of a ''zero-sum game'' [33], the GAN-based network architecture can generate the target image better, with a shorter running time and a higher quality of the generated image. Although the current network model based on GAN has great advantages, GAN itself cannot make full use of the relationship between multi-scale features, leading to the fuzzy details of the generated image and further affecting the accuracy of gait recognition. The self-attention (SA) mechanism has been of wide concern since it was put forward because it not only considers global aspects but also focuses on key points and can effectively establish the global dependence between features. At present, the SA mechanism has achieved good performance in image super-resolution reconstruction [34] and image generation [35]. Zhang and Bao [36] introduced the SA mechanism module into the GAN to learn the correlation between features to improve the quality of generated pictures. Given that GAN itself is prone to pattern collapse during training [33], this model also has such problems, and the gait recognition accuracy of the model needs to be further improved. The two-channel training strategy [37] uses a variety of losses to constantly constrain the generator through two network architectures that transform the input image into the target image and then reconstruct it into the input image, so that the generated

images are diverse and clear, which can well eliminate the influence caused by pattern collapse.

To sum up, to reduce the influence of cross-view on gait recognition, a GAN-based cross-view gait recognition network model combining multi-scale feature residual structure and SA mechanism was proposed. The multi-scale feature residual module (MFRM) was integrated into the feature extraction part of the proposed model. The module used differently sized receptive fields to fully obtain the local and global feature information of the image and generated more representative image feature maps, laying the foundation for the subsequent generation of high-quality gait maps. The model also introduced the SA mechanism module to adjust the information of channel dimensions, which was helpful to capture the correlation between feature information and contributed to the clearer and richer quality details of the generated gait map. To avoid pattern collapse in training, the two-channel network training strategy was adopted. The proposed network model is more conducive to the identification of identity information, improves the accuracy of gait recognition, and obtains gait results closer to the true value. This network model also exhibits excellent generalization ability.

The remainder of this paper is as follows: the proposed network structure and model loss were introduced in Section II. The experimental details, experimental results, and ablation experiments were presented in Section III. The conclusions were given in Section IV.

II. PROPOSED NETWORK STRUCTURE

The overall structure of this network is shown in Fig. 1, including two generators G and a discriminator D. The MFRM was integrated into G to improve the multi-scale feature extraction ability of the feature extraction module. Then, the SA mechanism module was introduced into G and D to adjust the feature information of the channel dimension, establish the correlation between different feature information, optimize G and D, and enhance the identification ability of the model. This network could realize the gait transformation of any gait image under any viewing angle.

In Fig. 1, x_i and y_i denote the gait image and view label of the subject i, respectively. x_i^s and y_i^s stand for the gait image and view label of the subject i at the original viewing angle s. x_i^t and y_i^t represent the gait image and view label the subject i at the target viewing angle t. Therein, $i, j \in \{1, 2, \ldots, N_x\}$ and $s, t \in \{1, 2, \ldots, N_v\}$, where N_x stands for the total number of subjects in the dataset and N_v represents the total number of viewing angles in the dataset. The one-hot vector coding method was adopted for the view label y_i^s. Specifically, 1 was allocated to the location corresponding to the viewing angle of the gait image and 0 to other locations, and the view label served as the indicator for the target viewing angle of the generator.

The GAN-based network model was trained using the two-channel network training strategy to prevent GAN from pattern collapse, i.e., to avoid the consistency of generated images, which would lead to the reduction of diversity. Two channels refer to the source domain data stream channel and the target domain data stream channel. The core of the model included two generators G and a discriminator D. The generators generated images with specific meanings according to the input information and tried to fool the ''discriminator'', which was used to distinguish whether the input images were generated or real, and the two were optimized alternately until the results were optimal. Given the same training process of the two channels, an explanation was given based on the source domain data stream channel.

Two source gait images x_i^s and x_j^t were randomly extracted, and view labels y_i^s and y_j^t were fabricated through one-hot coding. In the source domain data stream channel, the source gait image x_i^s and the target view indicator y_j^t were first connected as the input of the generator G to generate a synthesized gait image $x_i^{s\prime}$, which can be expressed by a formula $G(x_i^s, y_j^t) \rightarrow x_i^{s\prime}$. Then, the synthesized gait image $x_i^{s\prime}$ and the source view label y_i^s were connected and fed into the generator G to generate a reconstructed gait image $x_i^{\tilde{s}}$, which can be expressed by the formula: $G(x_i^{s\prime}, y_i^s) \rightarrow x_i^{\tilde{s}}$. Finally, the discriminator D judged the authenticity of the source gait image x_i^s and the reconstructed gait image $x_i^{\tilde{s}}$ and the type of their view domain, and constrained the generator and discriminator using the adversarial loss L_{adv} and view classification loss L_{view}. To facilitate the generator to generate an image similar to the target gait image x_i^t in a short time, the pixel-level loss L_{pixle} was used to minimize the error between the synthesized gait image $x_i^{s\prime}$ and the target gait image x_i^t. The cyclic consistency loss L_{cycle} constrained the generator by comparing the similarity between the source gait image x_i^s and the reconstructed image $x_i^{\tilde{s}}$, ensuring that the generator only changed the viewing angle of the source gait image x_i^s without changing its identity information.

From the overall architecture of the two-channel network, the source gait image x_i^s in the source domain data stream channel had gone through the process from the viewing angle s to t and finally to s, that is $x_i^s \rightarrow x_i^{s\prime} \rightarrow x_i^{\tilde{s}}$. The source gait image x_j^t in the data stream channel of the target domain had gone through the process from the viewing angle t to s and finally to t, that is $x_j^t \rightarrow x_j^{t\prime} \rightarrow x_j^{\tilde{t}}$. This two-channel network architecture forced the generator G to only change the feature information related to the viewing angle in the input image, not only enabling G to generate high-quality images but also effectively avoiding the tendency of generated images to similar distribution, lack of diversity, and pattern collapse.

A. GENERATOR NETWORK

To fully extract and utilize the features of the input image, the MFRM and the SA mechanism were introduced into the generator. As shown in Fig. 2, the generator network structure mainly included three parts: the down-sampling area, SA module, and up-sampling area. The feature extraction module and the MFRM constituted the down-sampling area. The original image and label code were first sent to the

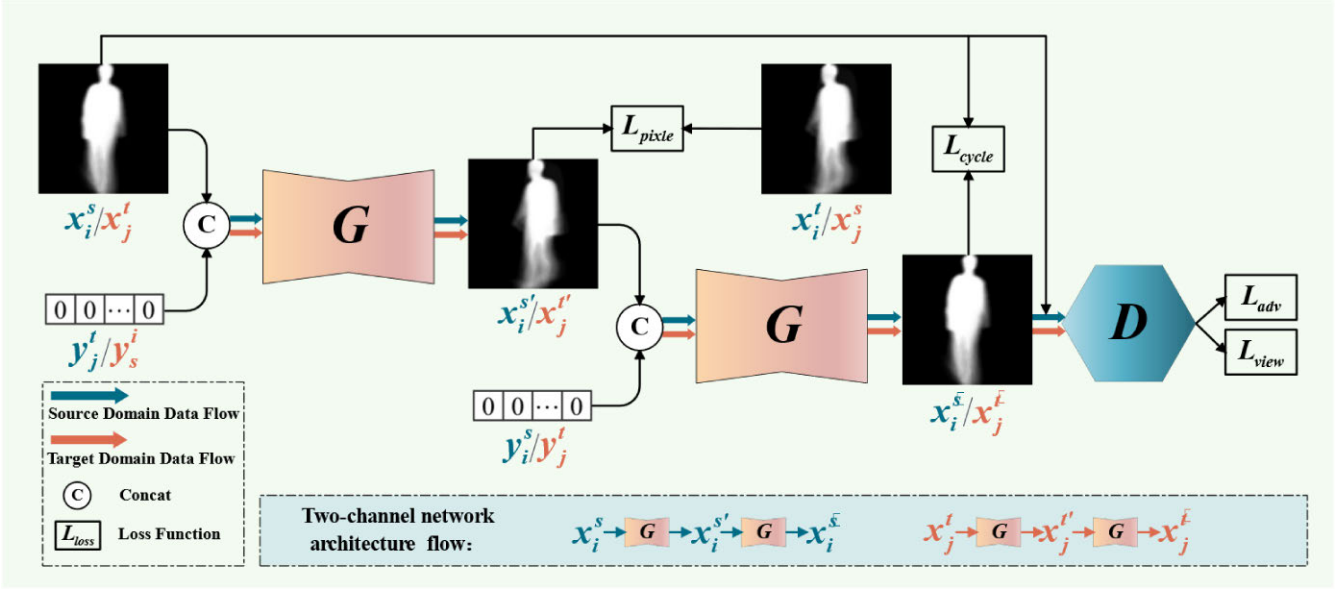

FIGURE 1. Overall structure of the network model.

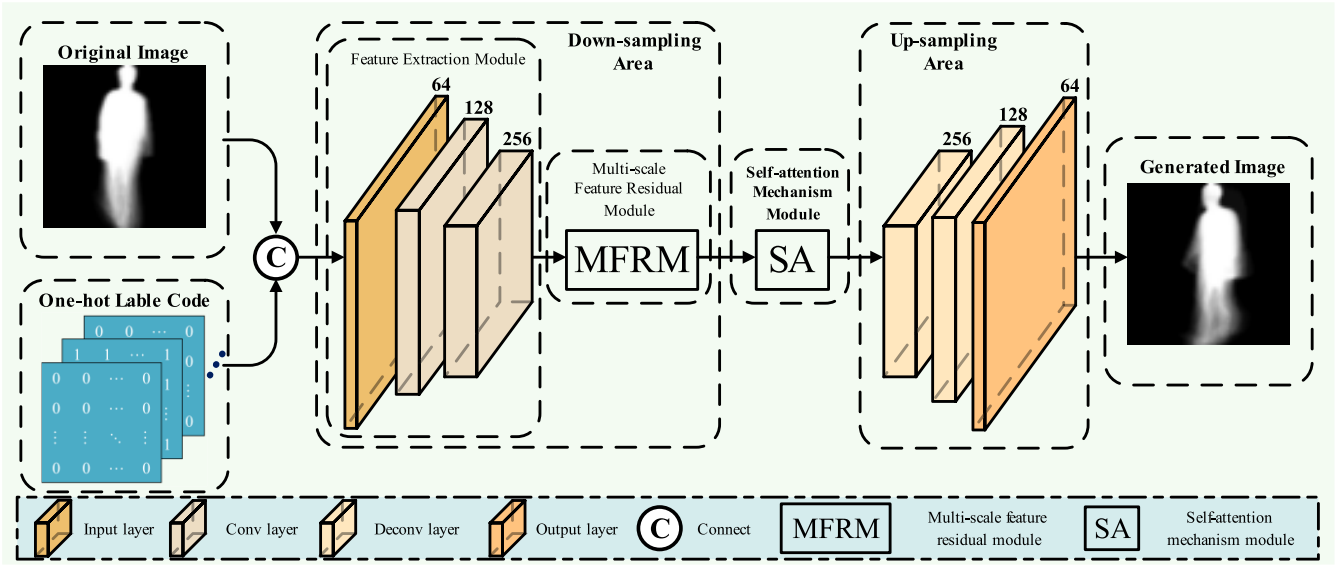

FIGURE 2. Generator network.

down-sampling area to extract the key features of multi-scale images. Then, the global dependency between the extracted features was established by the SA module. Finally, the up-sampling area generated an image with the same size as the original image and a transformed viewing angle based on the relevant feature information.

B. MULTI-SCALE FEATURE RESIDUAL MODULE

To better extract the global and local features of gait images, the MFRM was integrated into the feature extraction module of generator G to obtain more representative multi-scale feature information. As shown in Fig. 3, the structure consisted of main roads and branch roads. The main roads extracted important feature information from the input image silhouette with a large receptive field. The branch roads used small receptive fields to extract the local detail information of images. After extracting the feature information of the two main roads, two feature maps were obtained, the feature maps of the main roads and the branch roads were fused and output by using the residual, and the output feature map A was obtained. The combination of main and branch roads and residuals could not only improve the information extraction effect of the network but also reduce the calculation parameters of the network.

The parameter settings of the MFRM are shown in Table 1, which comprised three small modules. Among

them, K represents the size of the convolution kernel. C is the number of channels of the input tensor. S represents stride, and P is padding. Block1 is a convolution layer with a convolution kernel of 1 × 1. Block2 and Block3 were feature extraction modules for main roads and branch roads, respectively. The BN and LeakyReLU activation functions were used after each convolution layer.

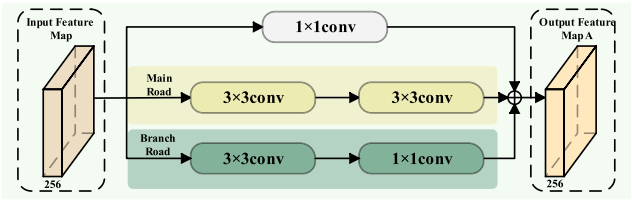

FIGURE 3. Multi-scale feature residual network.

TABLE 1. Parameter setting of multi-scale feature residual structure.

Layer Type	K	C	S	P	Normalization	Activation Function
Block1	1×1	256	1	0	BN	LeakyReLU
Block2	3×3	256	1	1	BN	LeakyReLU
	3×3	256	1	1	BN	LeakyReLU
Block3	3×3	256	1	1	BN	LeakyReLU
	1×1	256	1	0	BN	LeakyReLU

C. SELF-ATTENTION MECHANISM

The viewing angle refers to the angle between the camera and the traveling trajectory of the subject. Therefore, the gait energy maps (GEIs) synthesized under different viewing angles should be quite different, but the GEIs synthesized under 72°, 90°, and 108° were only slightly different in the details of the legs, as shown in Fig. 4. Among them, Fig. 4(a) represents the GEIs of the subject in three viewing angles under the normal walking condition, and Fig. 4(b) and 4(c) represent walking with a bag and walking with a coat respectively. Although the multi-scale feature residual structure could capture the global features and local features of gait images well by using differently sized convolution kernels, the generated images were blurred in detail and accompanied by virtual shadows because GAN could not make full use of all feature information. When the target viewing angle was 72°, the model could easily generate gait images with the viewing angle of 90° or 108° by mistake. The same phenomenon would happen when the target viewing angle was 90° or 108°. This is because the model only extracts the global information and local information of the image while not establishing the relationship between information, that is, obtaining the relationship between all the location features of the image is impossible. The SA mechanism can establish the relationship between local features and global features in a large range and adjust the feature information in the channel dimension, which has the key performance of capturing the internal correlation of features. The SA mechanism will use the features of all locations to generate certain detail information of the image, making the generated picture more realistic. Therefore, the SA mechanism was introduced into the model, and by adding the SA mechanism module in different locations of the generator and the discriminator, the proposed feature information was integrated to improve the quality of model generation.

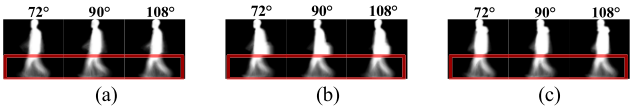

FIGURE 4. GEIs at a viewing angle of 72°–108° under three conditions. (a) GEIs at a viewing angle of 72°–108° under normal walking condition. (b) GEIs at a viewing angle of 72°–108° under walking with a bag condition. (c) GEIs at a viewing angle of 72°–108° under walking with a coat condition.

The structure of the SA mechanism module is shown in Fig. 5. First, the feature map $A \in R^{C\times H\times W}$ extracted from the previous layer was sent to two convolution layers with C/f output channels, and two new feature maps B and L were generated, respectively, where $\{B, L\} = R^{C\times H\times W}$ and $f \in (1, 2, 4, 8)$, and the value of f was taken as 8. The shape of feature maps B and L was adjusted as $R^{C\times N}$, where $N = H \times W$, which represents the number of features. Then, the transposed B and L were subjected to matrix multiplication and normalization through the Softmax function to calculate SA β, where $\beta \in R^{N\times N}$, as seen in Formula (1).

$$\beta_{uv} = \frac{exp(B_u \cdot L_v)}{\sum_{u=1}^{N} exp(B_u \cdot L_v)}, \tag{1}$$

where u and v are location subscripts; B_u is the feature ($u = 1, 2, \ldots, N$) of B at the location u; L_v represents the feature ($v = 1, 2, \ldots, N$) of L at the location v. $\beta_{uv} \in \beta$, which was used to measure the influence of the location u on the location v. If the features of the two locations were more similar, they were correlated to a greater degree.

The feature A was sent to a convolution layer with C output channels, and a new feature map $M \in R^{C\times H\times W}$ was generated, with its shape adjusted to $R^{C\times N}$. Then, the transposed M and β were subjected to matrix multiplication, and the resulting shape was adjusted as $R^{C\times H\times W}$. Finally, the result was multiplied by a scale parameter α and then subjected to element addition with the feature A to obtain the final output feature map $O \in R^{C\times H\times W}$, as seen in Formula (2).

$$O_v = \alpha \sum_{u=1}^{N} (\beta_{uv} M_u) + A_v, \tag{2}$$

where α is initialized 0 and gradually acquires a higher weight. The equation allows the inference that the feature map at each location is the weighted sum of all locations and the original feature. The finally generated features had global context information, and the context information was selectively aggregated according to SA. The SA

mechanism module endowed the representation of features with reciprocity through the information adjustment of channel dimensions, thus effectively establishing the global dependence between features.

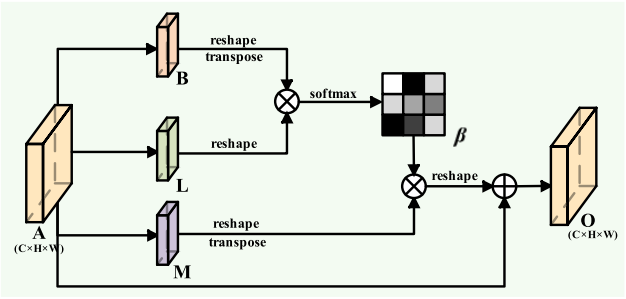

FIGURE 5. Self-attention mechanism.

D. DISCRIMINATOR NETWORK

To enhance the feature discrimination ability of the discriminator, the SA mechanism module was also integrated into the discriminator, as shown in Fig. 6. The discriminator network structure was mainly composed of a feature extraction module and two discrimination modules. The features extracted from the original image through the input layer and the downsampling layer were sent to the SA mechanism module first to adjust the information of the channel dimension. Then, such features entered the image authenticity judgment module D_x and the view judgment module D_y to judge the identity and viewing angle. When the input image came from the real sample distribution P_{data}, the output of D_x was 1; otherwise, it was 0. D_y mainly aimed to discriminate the viewing angle information of the input image, and its output was the coding value of the one-hot vector corresponding to its viewing angle information.

E. LOSS

To improve the accuracy of model recognition, the network was trained by combining the adversarial, view classification, pixel-level, and cyclic consistency loss.

(1) As the most important loss in GAN, the adversarial loss was used to constrain the generator and discriminator, and the objective function is shown in Formula (3).

$$\min_G \max_D L_{adv} = E_{x_i:P_{data}}[\log D_{adv}(x_i^t)] + E_{x_i:P_{data},y^t:P_{view}}[\log(1 - D_{adv}(G(x_i^s, y_j^t)))], \tag{3}$$

where $E[]$ represents the expected value of the distribution function, and the pedestrian gait images x_i^s and x_i^t followed the sample distribution P_{data}. The view label y_j^t obeyed the view indicator distribution P_{view}. $G(x_i^s, y_j^t)$ is the gait image generated after the source gait image x_i^s and target view indicator y_j^t were connected and sent into the generator G. The generator G aimed to minimize the objective function, while D_{adv} in the viewing angle discriminator should maximize the objective function.

(2) The view classification loss was used to measure the proximity between the viewing angle information of the gait image and the target view indicator. In general, the view classification loss was constructed using a cross-entropy function. When D was optimized, the following objective function should be minimized as shown in Formula (4).

$$L_{cls}^D = E_{x_i:P_{data}}[\log D_{cls}(x_i)], \tag{4}$$

where $D_{cls}(x_i)$ means that the input gait image comes from the real sample distribution P_{data}. When G was optimized, the input was the image under the target view indicator generated by the generator. By minimizing the objective function, the generator G was constrained to generate a gait image under the target view indicator, with its corresponding loss function shown in Formula (5).

$$L_{cls}^D = E_{x_i:P_{data},y^t:P_{view}}[\log D_{cls}(G(x_i^s, y_j^t))]. \tag{5}$$

(3) The pixel-level loss was used to minimize the error between the synthesized gait image $x_i^{s\prime}$ and the target gait image x_i^t. This enabled the generator to generate an image similar to the silhouette of the target gait map in a very short period, greatly shortening the "learning" period of unimportant feature information, with a relatively stable training environment. The loss is shown in Formula (6).

$$\min_G L_{pixel} = E_{x_i:P_{data},y^t:P_{view}}[\left\| G(x_i^s, y_j^t) - x_i^t \right\|_1], \tag{6}$$

where $\|\cdot\|_1$ represents the loss of pixel level L1, and x_i^s and x_i^t represent the source and target gait image, respectively.

(4) In this study, the generator was constrained by comparing the similarity between the reconstructed image $x_i^{\tilde{s}}$ and source image x_i^s using the cyclic consistency loss, as seen in Formula (7).

$$L_{cycle} = E_{x_i:P_{data},y^t:P_{view}} \left\| G(G(x_i^s, y_j^t), y_i^s) - x_i^s \right\|_1 + \left\| G(G(x_j^t, y_i^s), y_j^t) - x_j^t \right\|_1, \tag{7}$$

where $G(x_i^s, y_j^t)$ and $G(x_j^t, y_i^s)$ represent the synthesized gait images after the viewing angle is converted the first time by the generator, i.e., $x_i^{s\prime}$ and $x_j^{t\prime}$. $G(G\left(x_i^s, y_j^t\right), y_i^s)$ and $G(G\left(x_j^t, y_i^s\right), y_j^t)$ stand for the reconstructed gait images after the secondary viewing angle conversion by the generator, namely, $x_i^{\tilde{s}}$ and $x_j^{\tilde{t}}$. In this study, the quality of generated images was improved by minimizing the loss function.

The ultimate objective function of this model is the weighted sum of the above loss functions:

$$L_{all} = \lambda_1 L_{adv} + \lambda_2 L_{cls} + \lambda_3 L_{pixel} + \lambda_4 L_{cycle}, \tag{8}$$

where $\lambda_i (i = 1, 2, 3, 4)$ is the balance parameter between the losses. In this experiment, these parameters were set to $\lambda_1 = \lambda_2 = 1, \lambda_3 = 20, \lambda_4 = 10$ through constant adjustment, optimization [32], [36], and visualization of training images.

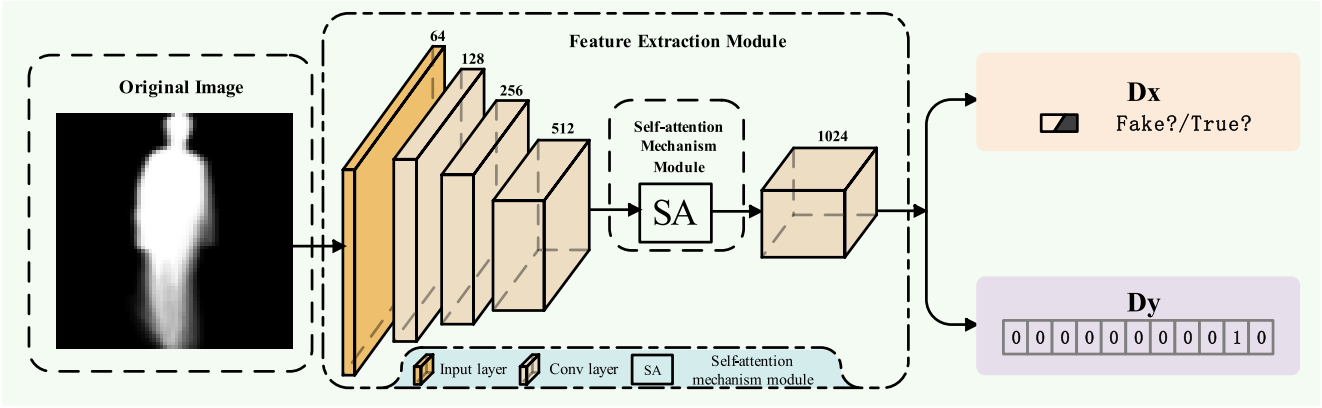

FIGURE 6. Discriminator network.

III. EXPERIMENT AND ANALYSIS

A. DATASET

In experiment was done with the proposed network model on the two public datasets CASIA-B and OU-MVLP, and the effectiveness of this network model was evaluated.

The CASIA-B [38] dataset is a multi-view gait recognition database provided by the Institute of Automation, Chinese Academy of Sciences, which has a wide range of viewing angles and is a commonly used gait dataset. This dataset contains the video sequences of 124 subjects under 3 walking states (normal walking, walking with a bag, and walking with a coat). The video sequences are gait sequences under 11 viewing angles (0°, 18°, …, 180°) for each subject, and 10 gait sequences are collected under each viewing angle, comprising 6 sequences (NM01-NM06) in normal walking conditions, 2 sequences (BG01-BG02) while carrying bags, and 2 sequences (CL01-CL02) while wearing coats. Therefore, each subject has 11× (6+2+2) =110 video sequences. The CASIA-B dataset was divided by the commonly used method [39], that is, the data of the first 62 pedestrians constituted the training dataset, and the data of the last 62 pedestrians constituted the test dataset. The test dataset was further divided into a gallery set and probe set according to different walking states, and the specific division is shown in Table 2.

TABLE 2. Experimental settings of CASIA-B.

Dataset	ID	Video Sequence
Train Set	001-062	NM01-NM06、BG01-BG02、CL01-CL02
Gallery Set	063-124	NM01-NM04
Probe Set	063-124	NM05-NM06、BG01-BG02、CL01-CL02

The OU-MVLP [40] dataset is a multi-view gait recognition database created by the Institute of Science and Industry of Osaka University in Japan. The OU-MVLP dataset contained 10,307 walking video sequences under 14 different viewing angles. Each video sequence contained the gait sequences of subjects under 14 viewing angles (0°, 15°, …,90°; 180°, 195°, …, 270°). Two gait sequences (#00–01) were collected under each viewing angle. The official division method [39] of the dataset was used, that is, the multi-view gait sequences of 5153 pedestrians constituted the training set and those of 5154 pedestrians formed the test set. In the testing stage, the #01 gait sequence was used as the gallery set and the #00 gait sequence as the probe set. The specific division is shown in Table 3.

TABLE 3. Experimental settings of OU-MVLP.

Dataset	ID	Video Sequence
Train Set	001-5153	#00 & #01
Gallery Set	5154-10307	#01
Probe Set	5154-10307	#00

B. EVALUATION INDEXES AND EXPERIMENTAL PARAMETERS

The Rank-1 index in the cumulative matching characteristic curve was used as the evaluation index for model recognition accuracy. First, the gait image in the probe set was converted to the target viewing angle corresponding to the gallery set, and the gait image $x_i^{s'}$ after view transformation was obtained. Then, the Nearest Neighbor Classifier was used to calculate the Euclidean distance between gait images in $x_i^{s'}$ and gallery set and then determine its identity information. Finally, whether the two images had the same identity was judged, and the Rank-1 recognition rate was obtained, which was positively correlated with the model recognition accuracy.

In the experiment, the network model was built by using the PyTorch framework and trained through an NVIDIA RTX 4090 graphics card. The size of the input and output GEIs of the network was 64 × 64 pixels. Batch_size was set to 64 [36]. The strategy of alternating iterative training [28] was adopted. After the discriminator D was trained 10 times, the generator G was updated once. In the process of training, the weights of all network models were randomly initialized by Gaussian distribution with a mean value of 0 and variance of 0.02. Network parameters were updated using the Adam optimizer, where $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and the initial

TABLE 4. Comparison between different methods in average Rank-1 recognition accuracy on the CASIA-B dataset. The highest score is marked in bold. All The scores are described in percentage (%). The red numbers in the table indicate the sorted results of various methods.

Methods	Probe Sample Set			Average
	NM05,NM06	BG01,BG02	CL01,CL02	
AE	59.3	37.2	24.2	40.2
GaitGAN	57.2	35.6	29.2	40.7
MGAN	68.1	54.7	31.5	51.4
GaitSet	92.0	84.3	62.5	79.6
Multi-branch	92.5	85.3	64.2	80.7
Two-stream	95.2	89.0	76.8	87.0
MT3D	94.4	89.8	75.6	88.6
GaitGL	95.9[3]	92.1[2]	78.3[3]	88.8[3]
MetaGait	**96.8**[1]	**94.0**[1]	83.5[2]	**91.4**[1]
Ours	96.0[2]	92.0[3]	**85.3**[1]	91.1[2]

learning rate was set to 0.0002 [24], [30], [36]. The model was trained for 200 K iterations. The learning rate remained unchanged in the first 100 K iterations, and the step strategy was adopted in the remaining 100 K iterations. The learning rate declined to 1% of the original per 1 K iterations until reaching 0. In the process of testing, the K value of the nearest classifier was taken as 5.

C. EXPERIMENTAL RESULTS OF CASIA-B AND OU-MVLP DATASETS

To verify the effectiveness of the proposed network model, it was compared with the latest methods such as Two-stream [23], Multi-branch [30], GaitSet [39], AE [41], GaitGAN [42], MGAN [43], MT3D [44], GaitGL [45] and MetaGait [46] on the CASIA-B dataset. The Rank-1 recognition accuracy of each method is listed in Table 4. The data in Table 4 are the average values of Rank-1 recognition accuracy under 11 viewing angles.

According to Table 4, in the process of recognizing different walking conditions by the same method, the recognition accuracy was the highest in the normal walking condition and the lowest in the coat-wearing condition. This is because the gait information of the subject is not blocked in the normal walking condition. However, the bag-carrying and coat-wearing conditions can block the gait information of the subject to a certain extent, where the latter will affect the gait information of the subject in a large range, increasing the difficulty in gait recognition and leading to a great decrease in the accuracy of gait recognition under the coat-wearing state.

For the accuracy recognition results of different methods in the same walking condition in Table 4, the network model in this research achieved good recognition accuracy in three conditions. The recognition accuracy was 85.3% under the condition of walking with a coat greatly affected by block, indicating that the network structure in this research could effectively overcome the influence of block on the gait recognition accuracy. This method performed better than other methods in both single-state recognition accuracy and the average value under three conditions. Compared with the MetaGait model, the accuracy of our method decreased by 0.8% and 2% in normal and backpack conditions, respectively. However, it was increased by 1.8% in the coat-wearing condition. When extracting the global dependence of various features, the MetaGait model not only applies the attention mechanism to the channel dimension, but also integrates the temporal and spatial dimensions. In addition, MetaGait model can adaptively capture the full-scale dependence of the space, channel, and time dimension. But our model only integrates the attention mechanism into the channel dimension. Therefore, compared with MetaGait model, the recognition accuracy of our model is slightly lower in some states. Nevertheless, the average Rank-1 recognition accuracy of our method also reached 91.1%, which is only 0.3% lower than the MetaGait model. This manifested as follows: by integrating the multi-scale feature residual structure and SA mechanism into the generator, the association between deep and shallow features could be effectively established, and the gait images with difficulty in identification were converted into images easy for identification, thus substantially enhancing the robustness and accuracy of the model.

Table 5 shows Rank-1 accuracy comparison results of different methods under the 11 validation views (excluding the same view) in the CASIA-B dataset. Fig. 7 shows the gait recognition performance of the proposed network model under three conditions. It can be seen from Table 5 and Fig. 7 that the model exhibited high accuracy under most viewing angles and conditions, but the Rank-1 accuracy near the viewing angle of 90° and 180° fluctuated considerably. This is because the gait image with a viewing angle of 90° is taken by the camera from the front side of the subject. Moreover, the gait images at 0° and 180° are taken by the camera from the front and back of the subject. Therefore, the GEI at 90° contained rich information about the movement characteristics of limbs. The GEI at 0° and 180° reflected the body and shape characteristics of the subject more. So, when the GEI at 90° was converted into the GEI at other angles, there was a gap between the newly generated GEI and its corresponding real image due to the lack of the body and shape characteristic information of the subject. The GEI at 90° was very similar to that at 72° and 108°, so the recognition accuracy at all angles (including 90°) near 90° increased, also explaining why the recognition accuracy at 180° decreased.

TABLE 5. Rank-1 accuracy (%) On CASIA-B Under 11 probe views excluding identical-view cases. The highest score is marked in bold. All the scores are described in percentage (%).

Probe	Method	0°-180°											Mean
		000°	018°	036°	054°	072°	090°	108°	126°	144°	162°	180°	
NM#5-6	AE	49.3	61.5	64.4	63.6	63.7	58.1	59.9	66.5	64.8	56.9	44.0	59.3
	MGAN	54.9	65.9	72.1	74.8	71.1	65.7	70.0	75.6	96.2	68.6	53.8	68.1
	GaitSet	86.8	95.2	98.0	94.5	91.5	89.1	91.1	95.0	97.4	93.7	80.2	92.0
	GaitGL	93.9	97.6	98.8	97.3	95.2	**92.7**	95.6	**98.1**	**98.5**	96.5	91.2	95.9
	Ours	**96.5**	**99.2**	**100.0**	**100.0**	**98.3**	77.8	**97.6**	96.8	97.8	**100.0**	**92.3**	**96.0**
BG#1-2	AE	29.8	37.7	39.2	40.5	43.8	37.5	43.0	42.	36.3	30.6	28.5	37.2
	MGAN	48.5	58.5	59.7	58.0	53.7	49.8	54.0	51.3	59.5	55.9	43.1	54.7
	GaitSet	79.9	89.8	91.2	86.7	81.6	76.7	81.0	88.2	90.3	88.5	73.0	84.3
	GaitGL	88.5	95.1	95.9	94.2	**91.5**	**85.4**	89.0	**95.4**	**97.4**	94.3	86.3	**92.1**
	Ours	**95.1**	**96.8**	**99.0**	**100.0**	87.7	63.4	**92.4**	95.3	95.9	**97.6**	**88.3**	92.0
CL#1-2	AE	18.7	21.0	25.0	25.1	25.0	26.3	28.7	30.0	23.6	23.4	19.0	24.2
	MGAN	23.1	34.5	36.3	33.3	32.9	32.7	34.2	37.6	33.7	26.7	21.0	31.5
	GaitSet	52.0	66.0	72.8	69.3	63.1	61.2	63.5	66.5	67.5	60.0	45.9	62.5
	GaitGL	70.7	83.2	87.1	84.7	**78.2**	**71.3**	78.0	83.7	83.6	77.1	63.1	78.3
	Ours	**93.2**	**97.4**	**95.2**	**98.0**	77.7	35.2	**82.0**	**87.3**	**97.4**	**95.2**	**79.9**	**85.3**

TABLE 6. Comparison between different methods in average Rank-1 recognition accuracy on the OU-MVLP dataset at four typical viewing angles. The highest score is marked in bold. All the scores are described in percentage (%).

Methods	Probe Set Views				Average
	0°	30°	60°	90°	
GEINet	8.2	32.3	33.6	28.5	25.7
TCC-GAN	21.0	38.0	38.0	31.0	32.0
3in+2diff	25.5	50.0	45.3	40.6	40.4
GaitSet	77.7	86.9	85.3	83.5	83.4
Multi-branch	71.7	86.1	82.3	83.6	80.9
GaitGL	84.9	91.1	91.1	90.3	89.4
GaitGCI	91.2	92.6	93.0	92.1	92.2
MetaGait	88.5	93.4	93.8	93.3	92.3
Ours	**94.5**	**99.1**	**99.2**	**98.2**	**97.8**

To prove its good generalization ability, the performance of the proposed network model was further evaluated on the OU-MVLP dataset. The gait sequences at four typical viewing angles (0°, 30°, 60°, 90°) were selected for evaluation according to the selection strategy of viewing angles proposed by Noriko et al. [40] in probe sets. The network model in this research was compared with eight methods—Multi-branch [30], TCC-GAN [32], GaitSet [39], GaitGL [45], MetaGait [46], GEINet [47], 3in+2diff [48] and GaitGCI [49]—and the comparison results are shown in Table 6. The table shows that the model had high recognition accuracy at four typical viewing angles, and the average recognition accuracy reached 97.8%, which was much higher than that obtained by other methods. A comparison of the evaluation results of the proposed network model on the CASIA-B dataset shows that the accuracy of the model was much higher than that in the CASIA-B dataset at several viewing angles. The possible reason is that there are more diverse and richer observed objects in the OU-MVLP dataset, and the model can learn more features that are convenient for identity recognition, so the overall recognition rate has been greatly improved.

The visualization results of this model on the CASIA-B dataset (left) and OU-MVLP dataset (right) are shown in Fig. 8. In the diagram, the red box (the first line) represents the input GEI randomly extracted from the probe set, the green box (the third line) indicates the real target GEI in the gallery set, and the middle line stands for the GEI generated by this model. Fig. 8 shows that the proposed model can generate a gait image that is highly similar to that under the real target viewing angle even in the case of a wide blocking range and a large change in the viewing angle.

D. ABLATION EXPERIMENT OF CASIA-B DATASET

The corresponding ablation experiment was performed to prove that the gait recognition accuracy could be improved greatly by the multi-scale feature residual structure and SA mechanism.

The SA mechanism can better capture the relationship between the local features and global features of images by adjusting the information of channel dimensions, so it was added to different locations of the model (generator and discriminator). In the experiment, it was first determined that the SA mechanism harvested the best recognition accuracy at the layer of the generator, and then it was determined that the SA mechanism reached the best recognition accuracy at the layer of the discriminator. Finally, the effect of the multi-scale feature residual structure was verified on the model added with the SA mechanism. Given that the redundancy of comparative experiments could be better avoided by first determining the location of the SA mechanism in the generator and discriminator, the location of the SA

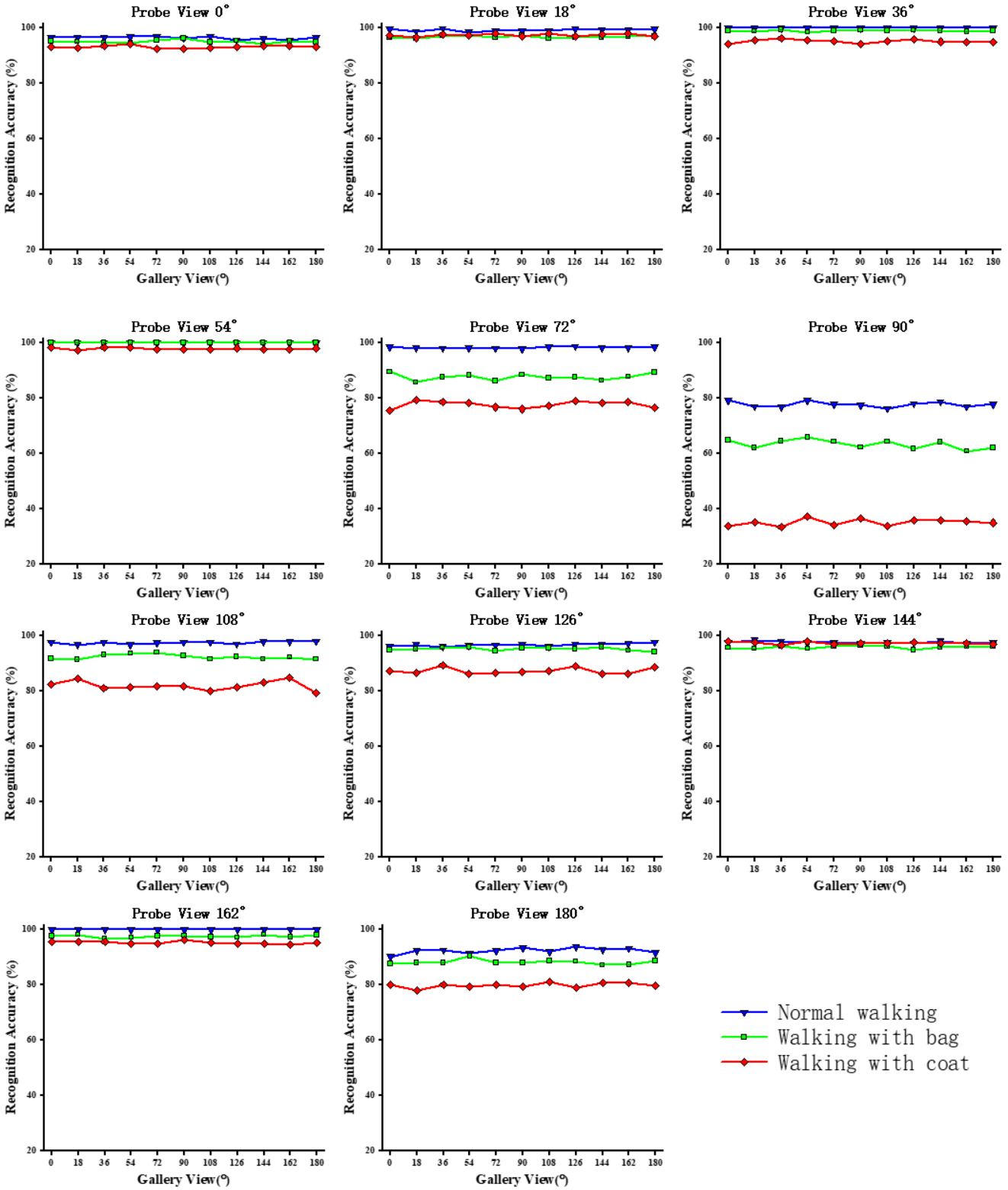

FIGURE 7. Recognition rate of three probe sets at different viewing angles on the CASIA-B dataset.

mechanism module was determined first, followed by the effectiveness validation of the multi-scale feature residual module.

The influence of each module on the experimental results was discussed on the CASIA-B dataset, and several groups of experiments were designed for the comparative analysis.

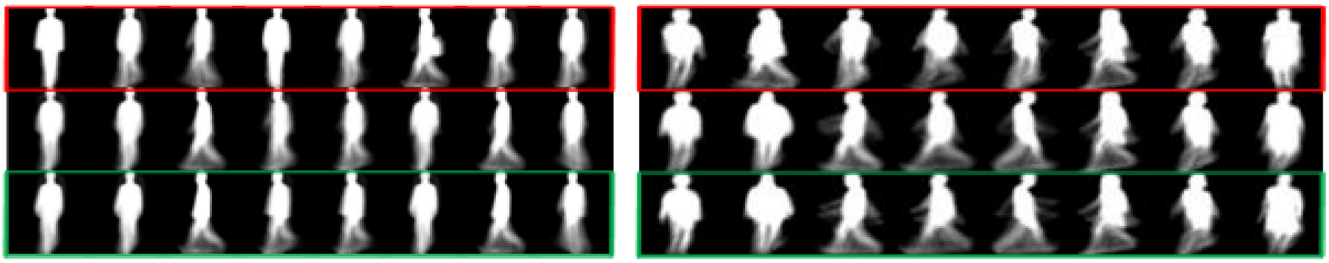

FIGURE 8. Visualization results under CASIA-B and Fig. 1. OU-MVLP datasets. (a) CASIA-B dataset. (b) OU-MVLP dataset.

The experimental parameters of each experiment are listed in Table 7. The experimental settings of each module were divided into three circumstances: (1) the influence of different locations of the SA mechanism module in the generator G on the experimental results (Table 8), (2) the influence of different locations of the SA mechanism module in the discriminator D on the experimental results (Table 9), and (3) the influence of MFRM on the experimental results (Table 10). Modules 1, 2, and 3 used in Table 7 are shown in Fig. 9.

TABLE 7. Multi-group experimental parameter settings. G (No SA) means that the generator G does not use the Sa mechanism module, G(SA+Deconv1) means that the SA mechanism module is placed before the deconvolution Layer 1 of the generator G, and D(Conv3+SA) means that the sa mechanism module is placed after the convolution Layer 3 of the discriminator D, and so on.

Experimental Number	Experimental Parameter Setting
Experiment 1	G(No SA)+D(No SA)+Module1
Experiment 2	G(SA+Deconv1)+D(No SA)+ Module1
Experiment 3	G(SA+Deconv2)+D(No SA)+ Module1
Experiment 4	G(SA+Deconv1)+D(Conv1+SA)+ Module1
Experiment 5	G(SA+Deconv1)+D(Conv2+SA)+ Module1
Experiment 6	G(SA+Deconv1)+D(Conv3+SA)+ Module1
Experiment 7	G(SA+Deconv1)+D(Conv3+SA)+ Module2
Experiment 8	G(SA+Deconv1)+D(Conv3+SA)+ Module3

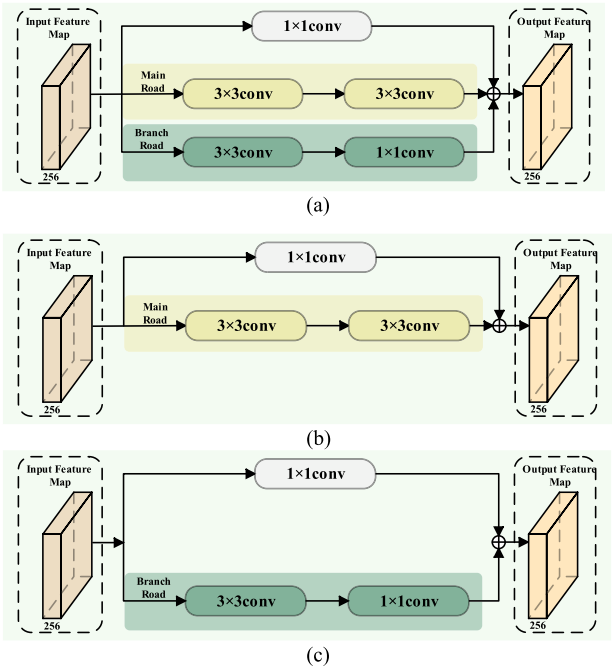

FIGURE 9. Comparison modules of the multi-scale feature residual structure. (a) Module 1. (b) Module 2. (c) Module 3.

Table 8 shows that the recognition effect was the best when the SA module was placed in front of the deconvolution layer 1 in the up-sampling area of the generator G, which was improved to some extent compared with the average recognition accuracy without using the SA module and placing it at other locations. Compared with Experiments 1 and 3, Experiment 2 reached the highest recognition accuracy, reaching 73.2%, under the coat-wearing condition (CL). This manifested in that adding an SA mechanism module before the deconvolution layer 1 of generator G could improve the recognition ability of the model under blocking conditions. This is because the SA mechanism could effectively establish the relationship between local features and global features extracted by the residual module and promote the generator to better generate images under the target viewing angle.

Table 9 shows that the recognition effect was the best when the SA module was placed after the convolution layer 3 of the feature extraction module of the discriminator D. Compared with other comparative experiments, the recognition accuracy of Experiment 6 was greatly improved in the bag-carrying condition (BG) and the coat-wearing condition (CL). The average recognition accuracy of Experiment 6 was 3.5% higher than that of Experiment 2.

To better prove the influence of the multi-scale feature residual structure on the experiment, the comparative experiment of the following three modules was designed, as shown in Fig. 9. Module 1 in Fig. 9(a) represents the proposed method, (b) Module 2 only keeps two 3×3 convolutions of the main road, and (c) Module 3 retains a 3×3 convolution and a 1×1 convolution of the branch road. The results of recognition accuracy under different modules are shown in Table 10.

As revealed in Table 10, the multi-scale feature residual result was the best under Module 1. Especially under the

TABLE 8. Influence of the location of the SA module at the generator on the recognition rate. The highest score is marked in bold. All the scores are described in percentage (%).

Experimental Number	Network model	Probe Sample Set			Average
		NM05,NM06	BG01,BG02	CL01,CL02	
Experiment 1	G(No SA)+D(No SA)+Module 1	97.7	90.4	72.8	87.0
Experiment 2	G(SA+Deconv1)+D(No SA)+module 1	**98.5**	**91.0**	**73.2**	**87.6**
Experiment 3	G(SA+Deconv2)+D(No SA)+module 1	97.9	90.0	71.0	86.3

TABLE 9. Influence of different locations of the SA mechanism at the discriminator on the recognition rate. The highest score is marked in bold. All the scores are described in percentage (%).

Experimental Number	Network Model	Probe Sample Set			Average
		NM05,NM06	BG01,BG02	CL01,CL02	
Experiment 2	G(SA+Deconv1)+D(No SA)+Module 1	**98.5**	91.0	73.2	87.6
Experiment 4	G(SA+Deconv1)+D(Conv1+SA)+Module1	98.0	91.2	75.6	88.3
Experiment 5	G(SA+Deconv1)+D(Conv2+SA)+module 1	98.1	91.0	73.9	87.7
Experiment 6	G(SA+Deconv1)+D(Conv3+SA)+Module 1	96.0	**92.0**	**85.3**	**91.1**

TABLE 10. Influence of different modules in the multi-scale feature residual structure on the recognition rate. The highest score is marked in bold. All the scores are described in percentage (%).

Experimental Number	Network Model	Probe Sample Set			Average
		NM05,NM06	BG01,BG02	CL01,CL02	
Experiment 6	G(SA+Deconv1)+D(Conv3+SA)+Module1	96.0	**92.0**	**85.3**	**91.1**
Experiment 7	G(SA+Deconv1)+D(Conv3+SA)+Module2	**98.1**	91.0	78.0	89.0
Experiment 8	G(SA+Deconv1)+D(Conv3+SA)+Module 3	94.3	89.9	83.9	89.4

walking with a coat (CL), the recognition accuracy of Experiment 6 was as high as 85.3%, which was better than that of Experiments 7 and 8. The average recognition accuracy of Experiment 6 was 2% higher than that of Experiments 7 and 8, proving that the fusion of the residual block with differently sized receptive fields could improve the feature extraction ability of the model and enhance the gait recognition accuracy under cross-view circumstances.

IV. CONCLUSION

To solve the low accuracy of gait recognition under cross-view conditions, a cross-view gait recognition network model combining the MFRM and SA mechanism was proposed. Specifically, the multi-scale feature residual structure was integrated into the feature extraction module of the generator to fully extract the deep features and shallow features of the input gait image. The SA mechanism module was used to adjust the information of the channel dimension of the extracted multi-scale features and establish the global dependence between the feature information. In addition, the generator was constrained using the training strategy of the two-channel network so that the feature distribution of the generated image was extremely similar to the target image, thereby improving the quality of the generated image. The experimental results on CASIA-B and OU-MVLP datasets show that the proposed method is superior to the commonly used algorithms. The ablation experiment on the CASIA-B dataset also proves the effectiveness of the MFRM and SA mechanism module. Given the large oscillation of the proposed method in the CASIA-B probe set at viewing angles of 90° and 180°, the multi-view mean will be adopted. Here, three gait images are generated simultaneously with the target viewing angle as the mean, and the mean recognition rate of the three images is taken as the recognition rate at the target viewing angle, thereby enhancing the model robustness under a single viewing angle and strengthen its generalization ability for different datasets.

ACKNOWLEDGMENT

The authors sincerely thank the CASIA-B Dataset provided by the Institute of Automation, Chinese Academy of Sciences http://www.cbsr.ia.ac.cn/china/Gait%20Databases%20CH .asp, and the OU-MVLP Dataset provided by the Institute of Science and Industry, Osaka University, Japan http://www.am.sanken.osaka-u.ac.jp/BiometricDB/GaitMVLP.html. They are also very grateful for the valuable comments and contributions of the anonymous reviewers and the members of the editorial team.

REFERENCES

[1] N. Sadeghzadehyazdi, T. Batabyal, and S. T. Acton, "Modeling spatiotemporal patterns of gait anomaly with a CNN-LSTM deep neural network," *Expert Syst. Appl.*, vol. 185, Dec. 2021, Art. no. 115582, doi: 10.1016/j.eswa.2021.115582.

[2] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Cross-view gait recognition using pairwise spatial transformer networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 260–274, Jan. 2021, doi: 10.1109/TCSVT.2020.2975671.

[3] O. Elharrouss, N. Almaadeed, S. Al-Maadeed, and A. Bouridane, "Gait recognition for person re-identification," *J. Supercomput.*, vol. 77, no. 4, pp. 3653–3672, Apr. 2021, doi: 10.1007/s11227-020-03409-5.

[4] P. Chattopadhyay, S. Sural, and J. Mukherjee, "Frontal gait recognition from incomplete sequences using RGB-D camera," *IEEE Trans. Inf. Forensics Security*, vol. 9, no. 11, pp. 1843–1856, Nov. 2014, doi: 10.1109/TIFS.2014.2352114.

[5] I. Bouchrika, "A survey of using biometrics for smart visual surveillance: Gait recognition," in *Surveillance in Action* (Advanced Sciences and Technologies for Security Applications), P. Karampelas and T. Bourlai, Eds. Cham, Switzerland: Springer, 2018, pp. 3–23, doi: 10.1007/978-3-319-68533-5_1.

[6] Y. Wang. (2017). *View-Invariant Gait Recognition Based on Kinect Skeleton Information*. Shandong University. Accessed: Apr. 5, 2023. [Online]. Available: https://kns.cnki.net/kcms2/article/abstract?v=3uoqIhG8C475KOm_zrgu4lQARvep2SAkVtq-vp-8Qbjqyhl E-4l1YtDCFxigzuS7liUu5gLnPU8d5rNqVdEiDRS-L191aNZY& uniplatform=NZKPT

[7] X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, "Gait energy response function for clothing-invariant gait recognition," in *Computer Vision—ACCV 2016* (Lecture Notes in Computer Science), vol. 10112, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 257–272, doi: 10.1007/978-3-319-54184-6_16.

[8] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, "Speed invariance vs. stability: Cross-speed gait recognition using single-support gait energy image," in *Computer Vision—ACCV 2016* (Lecture Notes in Computer Science), vol. 10112, S.-H. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds. Cham, Switzerland: Springer, 2017, pp. 52–67, doi: 10.1007/978-3-319-54184-6_4.

[9] M. Tariq and M. A. Shah, "Review of model-free gait recognition in biometric systems," in *Proc. 23rd Int. Conf. Automat. Comput. (ICAC)*, 2017, pp. 1–7.

[10] J. Zhang and Y. He, "Deep learning for gait recognition: A survey," *Pattern Recognit. Artif. Intell.*, vol. 31, no. 5, pp. 442–452, 2018, doi: 10.16451/j.cnki.issn1003-6059.201805006.

[11] W. Xing, Y. Li, and S. Zhang, "View-invariant gait recognition method by three-dimensional convolutional neural network," *J. Electron. Imag.*, vol. 27, no. 1, p. 1, Jan. 2018, doi: 10.1117/1.JEI.27.1.013010.

[12] Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, "A comprehensive study on cross-view gait based human identification with deep CNNs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 2, pp. 209–226, Feb. 2017, doi: 10.1109/TPAMI.2016.2545669.

[13] M. B. Hasan, T. Ahmed, S. Ahmed, and M. H. Kabir, "GaitGCN++: Improving GCN-based gait recognition with part-wise attention and DropGraph," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 7, Jul. 2023, Art. no. 101641, doi: 10.1016/j.jksuci.2023.101641.

[14] R. Liao, S. Yu, W. An, and Y. Huang, "A model-based gait recognition method with body pose and human prior knowledge," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107069, doi: 10.1016/j.patcog.2019.107069.

[15] D. Thapar, G. Jaswal, A. Nigam, and C. Arora, "Gait metric learning Siamese network exploiting dual of spatio-temporal 3D-CNN intra and LSTM based inter gait-cycle-segment features," *Pattern Recognit. Lett.*, vol. 125, pp. 646–653, Jul. 2019, doi: 10.1016/j.patrec.2019.07.008.

[16] J. Han and B. Bhanu, "Individual recognition using gait energy image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 2, pp. 316–322, Feb. 2006, doi: 10.1109/TPAMI.2006.38.

[17] A. Al-Tayyan, K. Assaleh, and T. Shanableh, "Decision-level fusion for single-view gait recognition with various carrying and clothing conditions," *Image Vis. Comput.*, vol. 61, pp. 54–69, May 2017, doi: 10.1016/j.imavis.2017.02.004.

[18] K. Bashir, T. Xiang, and S. Gong, "Gait recognition using gait entropy image," in *Proc. 3rd Int. Conf. Imag. Crime Detection Prevention (ICDP)*, London, U.K., 2009, pp. 1–2, doi: 10.1049/ic.2009.0230.

[19] T. H. W. Lam, K. H. Cheung, and J. N. K. Liu, "Gait flow image: A silhouette-based gait representation for human identification," *Pattern Recognit.*, vol. 44, no. 4, pp. 973–987, Apr. 2011, doi: 10.1016/j.patcog.2010.10.011.

[20] W. Kusakunniran, Q. Wu, J. Zhang, and H. Li, "Gait recognition under various viewing angles based on correlated motion regression," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 6, pp. 966–980, Jun. 2012, doi: 10.1109/TCSVT.2012.2186744.

[21] D. Muramatsu, Y. Makihara, and Y. Yagi, "View transformation model incorporating quality measures for cross-view gait recognition," *IEEE Trans. Cybern.*, vol. 46, no. 7, pp. 1602–1615, Jul. 2016, doi: 10.1109/TCYB.2015.2452577.

[22] S. Xu, F. Zheng, J. Tang, and W. Bao, "Dual branch feature fusion network based gait recognition algorithm," *J. Image Graph.*, vol. 27, no. 7, pp. 2263–2273, 2022.

[23] J. Zhang, J. Li, and H. Gan, "Gait recognition combined with two-stream network and pyramid mapping," *Appl. Res. Comput.*, vol. 39, no. 6, pp. 1911–1915, 2022, doi: 10.19734/j.issn.1001-3695.2021.11.0636.

[24] H. Zhang and P. Tian, "Gait recognition method combining residual network and multi-level block structure," *J. Electron. Meas. Instrum.*, vol. 36, no. 6, pp. 66–72, 2022, doi: 10.13382/j.jemi.B2104954.

[25] K. Wang, Y. Lei, and J. Zhang, "Two-stream gait network for cross-view gait recognition," *Pattern Recognit. Artif. Intell.*, vol. 33, no. 5, pp. 383–392, 2020, doi: 10.16451/j.cnki.issn1003-6059.202005001.

[26] Y. Cui and Y. Kang, "Multi-modal gait recognition via effective spatial–temporal feature fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, BC, Canada, Jun. 2023, pp. 17949–17957, doi: 10.1109/CVPR52729.2023.01721.

[27] H. Zhu, Z. Zheng, and R. Nevatia, "Gait recognition using 3-D human body shape inference," 2022, *arXiv:2212.09042*.

[28] E. Kropat, A. Özmen, G.-W. Weber, S. Meyer-Nieberg, and O. Defterli, "Fuzzy prediction strategies for gene-environment networks—Fuzzy regression analysis for two-modal regulatory systems," *RAIRO-Oper. Res.*, vol. 50, no. 2, pp. 413–435, Apr. 2016, doi: 10.1051/ro/2015044.

[29] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, Dec. 2000, doi: 10.1109/34.895972.

[30] S. Hu, X. Wang, and Y. Liu, "Cross-view gait recognition method based on multi-branch residual deep network," *Pattern Recognit. Artif. Intell.*, vol. 34, no. 5, pp. 455–462, 2021, doi: 10.16451/j.cnki.issn1003-6059.202105008.

[31] X. Zhai, "Research on cross-view gait recognition algorithms based on deep learning," Shandong Univ., Jinan, China, Tech. Rep., 2020, doi: 10.27272/d.cnki.gshdu.2020.004901.

[32] Y. Wang and Y. Xia, "Cross-view gait recognition on two-channel cycle consistency GAN," *Appl. Res. Comput.*, vol. 39, no. 1, pp. 259–264, 2022, doi: 10.19734/j.issn.1001-3695.2021.05.0202.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020, doi: 10.1145/3422622.

[34] L. Bu, D. Dai, Z. Zhang, Y. Yang, and M. Deng, "Hyperspectral super-resolution reconstruction network based on hybrid convolution and spectral symmetry preservation," *Remote Sens.*, vol. 15, no. 13, p. 3225, Jun. 2023, doi: 10.3390/rs15133225.

[35] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2310–2314, Dec. 2017, doi: 10.1109/LGRS.2017.2762694.

[36] H. Zhang and W. Bao, "The cross-view gait recognition analysis based on generative adversarial networks derived of self-attention mechanism," *J. Image Graph.*, vol. 27, no. 4, pp. 1097–1109, 2022.

[37] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251, doi: 10.1109/iccv.2017.244.

[38] S. Yu, D. Tan, and T. Tan, "A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Jun. 2006, pp. 441–444, doi: 10.1109/ICPR.2006.67.

[39] H. Chao, Y. He, J. Zhang, and J. Feng, "GaitSet: Regarding gait as a set for cross-view gait recognition," in *Proc. AAAI*, Jul. 2019, vol. 33, no. 1, pp. 8126–8133, doi: 10.1609/aaai.v33i01.33018126.

[40] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition," *IPSJ Trans. Comput. Vis. Appl.*, vol. 10, no. 1, pp. 1–12, Dec. 2018, doi: 10.1186/s41074-018-0039-6.

[41] S. Yu, Q. Wang, L. Shen, and Y. Huang, "View invariant gait recognition using only one uniform model," in *Proc. 23rd Int. Conf. Pattern Recognit. (ICPR)*, Cancun, Mexico, Dec. 2016, pp. 889–894, doi: 10.1109/ICPR.2016.7899748.

[42] S. Yu, H. Chen, E. B. G. Reyes, and N. Poh, "GaitGAN: Invariant gait feature extraction using generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 532–539, doi: 10.1109/CVPRW.2017.80.

[43] Y. He, J. Zhang, H. Shan, and L. Wang, "Multi-task GANs for view-specific feature learning in gait recognition," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 1, pp. 102–113, Jan. 2019, doi: 10.1109/TIFS.2018.2844819.
[44] P. Lin, P. Sun, G. Cheng, S. Xie, X. Li, and J. Shi, "Graph-guided architecture search for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 4202–4211, doi: 10.1109/CVPR42600.2020.00426.
[45] B. Lin, S. Zhang, and X. Yu, "Gait recognition via effective global-local feature representation and local temporal aggregation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 14628–14636, doi: 10.1109/ICCV48922.2021.01438.
[46] H. Dou, P. Zhang, W. Su, Y. Yu, and X. Li, "MetaGait: Learning to learn an omni sample adaptive representation for gait recognition," 2023, *arXiv:2306.03445*.
[47] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "GEINet: View-invariant gait recognition using a convolutional neural network," in *Proc. Int. Conf. Biometrics (ICB)*, Halmstad, Sweden, Jun. 2016, pp. 1–8, doi: 10.1109/ICB.2016.7550060.
[48] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, "On input/output architectures for convolutional neural network-based cross-view gait recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 9, pp. 2708–2719, Sep. 2019, doi: 10.1109/TCSVT.2017.2760835.
[49] H. Dou, P. Zhang, W. Su, Y. Yu, Y. Lin, and X. Li, "GaitGCI: Generative counterfactual intervention for gait recognition," 2023, *arXiv:2306.03428*.

JINGXUE WANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in photogrammetry and remote sensing from Liaoning Technical University, Fuxin, Liaoning, China, in 2004, 2007, and 2011, respectively. From 2015 to 2021, she was a part-time Postdoctoral Fellow with Southwest Jiaotong University, Chengdu, Sichuan, China. From 2019 to 2020, she was a Visiting Scholar with the University of Calgary, Calgary, Canada. She has been with Liaoning Technical University, since 2011, where she is currently a Professor with the School of Geomatics. Her current research interests include image matching and point cloud data processing.

JUN GUO (Member, IEEE) received the B.Eng. degree in surveying and mapping engineering from the Changchun Institute of Technology, Changchun, Jilin, China, in 2021. He is currently pursuing the master's degree with the School of Geomatics, Liaoning Technical University.

ZHENGHUI XU (Member, IEEE) received the B.Eng. degree in surveying and mapping engineering from the Liaoning Institute of Science and Technology, Benxi, Liaoning, China, in 2018, and the M.Sc. degree from the School of Geomatics, Liaoning Technical University, Fuxin, China, in 2021, where he is currently pursuing the Ph.D. degree.