Gait Recognition Based on Temporal Gait Information Enhancing

Qizhen Chen[1,2,3,4], Xin Chen[1,2,3,4], Xiaoling Deng[1,2,3,4(✉)], and Yubin Lan[1,2,3,4]

[1] College of Electronic Engineering, College of Artificial Intelligence, South China Agricultural University, Guangzhou 510642, China
icefloor@stu.scau.edu.cn, {chenxin,dengxl,ylan}@scau.edu.cn
[2] National Center for International Collaboration Research on Precision Agricultural Aviation Pesticide Spraying Technology, Guangzhou 510642, China
[3] Guangdong Laboratory for Lingnan Modern Agriculture, Guangzhou 510642, China
[4] Guangdong Engineering Technology Research Center of Smart Agriculture, Guangzhou 510642, China

Abstract. Gait recognition is a long range biometric technology that identifies individuals by their walking patterns. Currently, gait recognition primarily extracts gait features using convolutional neural networks, which are based on either the global appearance or local human body regions. However, the global feature methods are lack of long range interactions in different local regions and lose temporal features by some extent, and the local feature method segmenting gait silhouettes into blocks limits the ability to characterize local feature weights. In this paper, we propose a gait recognition method that enhances interactions between local regions. To implement this method, we construct a new feature enhancement module, which is a global and local feature extractor based on SENet (GLFES), to enhance the recognition of local features using the attention mechanism. Extensive experiments based on our proposed method have been conducted on the public datasets CASIA-B and OUMVLP to achieve state-of-the-art performances.

Keywords: Gait recognition · Squeeze and excitation · Convolutional neural network · Human body alignment · Global and local features

1 Introduction

As a kind of unique biometric features, gait is long-range, non-contact and difficult to disguise. Moreover, gait samples can be obtained without subjects' cooperation. It has a very wide range of applications in security surveillance, criminal investigation surveillance and the public domain (Fig. 1).

Existing global or local feature extraction methods [1,10,12,17,26] mainly focus on extracting spatial features, while temporal features are limited by some extent. Specifically, the max-pooling operation in the temporal dimension easily

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 451–463, 2024.
https://doi.org/10.1007/978-3-031-53308-2_33

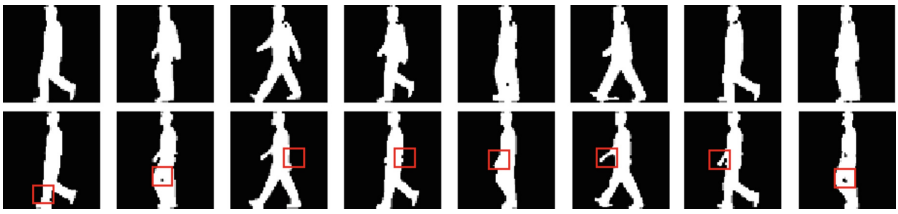
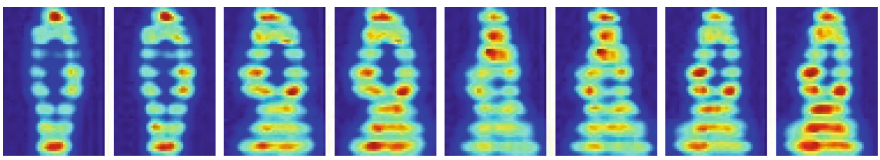


Fig. 1. The gait silhouettes images are from individuals 33 and 53, taken every 4 frames. It can be observed that subtle changes need to be slowly detected through timing, as a single silhouette image alone cannot depict them. Therefore, the temporal features of gait are crucial, especially when individuals are wearing coats.

loses many gait temporal features. Moreover, existing methods mainly segment gait features into chunks and extract local features of different local regions separately, the interacting relationships between different human body parts have not been discovered enough, which thus limits gait recognition accuracies.

To address these issues, in this paper, we propose a gait recognition method that enhances the temporal gait features and interactions between local features. Specifically, we build a new feature enhancement module, called Global and Local Feature Extractor based on SeNet (GLFES), to enhance the interactions between local features by integrating attention mechanism. This method is realized by a squeeze and excitation Module (SEM) in the network, a module that is capable of enhancing inter-regional interactions. We can see the effect of SEM in Fig. 2. We show visualization maps under different views, including 0°, 54°, 90° and 126°.

At the same time, we design a Temporal Global and Local Aggregator (TGLA) to extract temporal global and local features in a principled way. The global timing feature extractor focuses on the timing features of the entire gait sequence, while the local timing feature extractor splits the gait sequence in the time dimension, focusing on the gait details of adjacent frames. This then gives the model better recognition abilities by merging both global and local features.



(a) 0° (b) 54° (c) 90° (d) 126°

Fig. 2. The convolutional feature map visualization after SEM. The visualization on the left is without SEM module, and the visualization on the right is with SEM module.

Finally, we design a novel Temporal Feature Aggregator (TFA). The gait features have two dimensions, height and width, and the features in the higher dimension are more discriminative than the wider dimension. Therefore, by pooling the wide dimension and reducing the number of parameters of the gait features as input to the TFA, the recognition accuracy of the model can be improved. The highlighted contributions of our method are listed as follows:

(1) We propose a simple and lightweight application of convolution, called Temporal Global and Local Aggregator (TGLA), to facilitate refined learning of temporal features at the local level. The core idea of TGLA is to constraint the convolutional temporal perceptual wilderness and focus more on local timing information of gait-adjacent frames, then the local temporal feature are enhanced.
(2) We propose a novel local feature enhancement module (SEM) that maximises the usage of local features by interacting with local features in different regions.
(3) We propose a lightweight convolutional application, called Temporal Feature Aggregator (TFA), which is able to improve the comprehensive performance of the model.
(4) We conduct extensive experiments on the public datasets CASIA-B and OUMVLP. Experimental estimates show that the proposed method can achieve the state-of-the-art performance.

2 Related Work

Current gait recognition studies prove the importance of spatial feature extraction and modeling from time series [2–4,6,7,9,14,23,24,27,31,32]. In order to obtain more discriminative features from gait sequences, most of the existing models are based on CNNs, and they use 2D [2,32] or 3D [14,16,22–25] convolution for feature extraction along the spatial dimension with good success. The importance of different human body parts in gait recognition is different, and performing the same scanning operation for all gait sequences often overlooks this characteristic. To obtain more detailed information about the different human body parts, GaitSet [2,3], GaitPart [7], GLN [9], MT3D [15] tried to slice the output features into m blocks along the horizontal dimension, which can learn unique gait features of different body parts.

In addition, to better obtain discriminative gait features, many studies have integrated the entire gait sequence into one frame [14,29]. At the same time, there are many studies that extracted frame-level features from gait sequences by CNNs and applied a max-pooling operation on the temporal dimension [2,9], which easily limits the interrelationships and interactions between different gait frames.

In order to better obtain the relationships between consecutive gait frames, the original max-pooling operation is replaced by LSTM to integrate the gait features in the time series to generate the final gait features [5,13,18,32], and the

whole pipeline retains the non-essential order constraint in the gait sequences. These methods are good at extracting spatial and temporal features from gait sequences, ignore the spatio-temporal dependencies between non-local features.

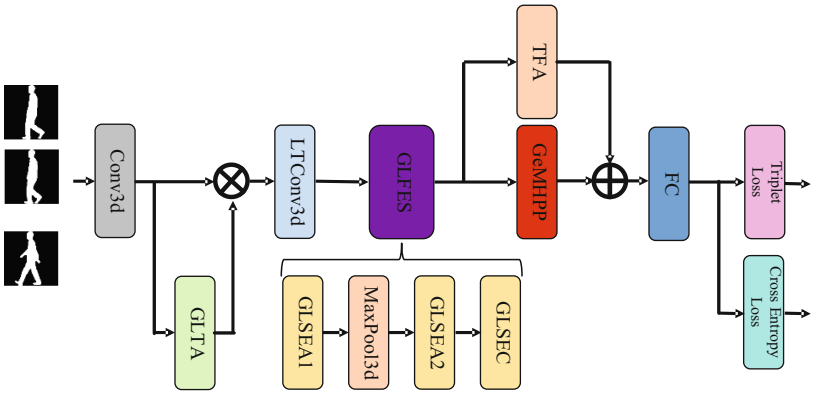


Fig. 3. The overview of the whole GaitSE framework. 'Conv3d' and 'LTConv3d' denote three-dimensional convolution. 'TGLA' represents the Temporal Global and Local Aggregator. 'GLFES' represents the Global and Local Feature Extrator based on SeNet. 'TFA' represents Temporal Feature Aggregator. 'GeMHPP' represents the Generalized-Mean Horizontal Pyramid pooling. 'FC' represents Fully Connected layer. 'Triplet Loss' and 'Cross Entropy Loss' represent two kinds of loss functions

3 Method

In this section, the pipeline of GaitSE is first described, then the Temporal Global and Local Aggregator (TGLA) is described, followed by the Global and Local Feature Extractor based on SeNet(GLFESN), ending with the Temporal Feature Aggregator (TFA) and implementation details. The overall framework is presented in Fig. 3.

3.1 Pipeline

To obtain more holistic gait features, we first extract shallow features from the original gait sequences by 3D CNNs. Next, the Temporal Global and Local Aggregator (TGLA) was designed to extract a combination of global and local temporal information. Later, the Local Temporal 3D Convolution (LTConv3d) is designed to replace the original Max-pooling operation to retain more spatio-temporal information to ensure more comprehensive temporal information. After that, the Global and Local Feature Extractor based on SeNet (GLFES) is designed to enhance global and local information. Then, we propose the Temporal Feature Aggregator (TFA) to integrate the global temporal information. Finally, the triplet loss and cross entropy loss are used as our loss functions to train the model.

3.2 Temporal Global and Local Aggregator

We propose a Temporal Global and Local Aggregator (TGLA). The TGLA module consists of two 3D CNNS, one for global temporal information and the other for local temporal information. Since global and local temporal information are considered at the same time, the gait features extracted by TGLA are comprehensive, as shown in Fig. 4.

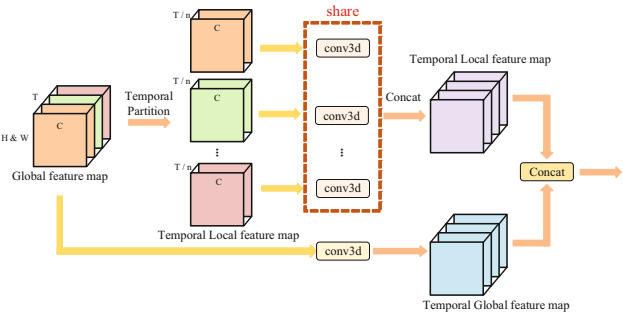


Fig. 4. Architectures of Temporal Global and Local Aggregator. 'H', 'W', 'C' and 'T' denote the height, width, number of channels and length of the gait sequence. 'n' denotes the number of cuts along the time dimension. 'Temporal Partition' denotes the segmentation of features along the time dimension and 'N' represents the number of segmented regions. 'Conv3d' is a 3D convolution operation, and 'share' denotes these 3D convolution shared parameters. 'Concat' indicates a concat operation in the temporal dimension

3.3 Global and Local Feature Extractor Based on SeNet

We propose a Global and Local Feature Extractor based on SeNet (GLFES). The first of this paragraph for the two forms of fusion methods, local features and global features, there are two classical methods, one is addition and the other is concat in this high dimension. The module based on additional fusion method we define as GLSEA, the module based on concat fusion method we define as GLSEC, the difference between the two lies in the final fusion method. The GLFES module consists of four layers 'GLSEA1-MaxPool3d-GLSEA2-GLSEC' as shown in Fig. 3.

The principle of TGLA has been shown above, and the difference between TGLA and GLSE lies in how the local feature map is partitioned and the Squeeze and Excitation Module (SEM). The segmentation of TGLA is along the time dimension, and GLSE is a horizontal segmentation. The SEM is shown in Fig. 5

Given $X_l \in \mathbb{R}^{H\times W\times T\times C_l}$ as the final local feature map. Therefore, the squeeze and excitation module can be formulated as:

$$Y_f = F_{se}(Reshape(Max(X_l))) \tag{1}$$

$$Y_{end} = F_{scale}(Reshape(Y_f), x_l) \tag{2}$$

where $Max(.)$ in maximizes the width of the feature, and $Reshape(.)$ denotes merging the height and time dimensions of the feature. $F_{se}(.)$ indicates that 'FC-ReLU-FC-Sigmoid' operations are performed, and and the output dimension of the first FC is $\frac{C_l}{r}$, and the output dimension of the second FC is C_l. $Y_f \in \mathbb{R}^{H\times 1\times T\times C_l}$ indicates the output of the Eq. 1. The $Reshape$ in Eq. 2 means to separate the dimensions. $F_{scale}(.,.)$ denotes width-wise multiplication between the feature map. $Y_{end} \in \mathbb{R}^{H\times W\times T\times C_l}$ denotes the output of the Eq. 2.

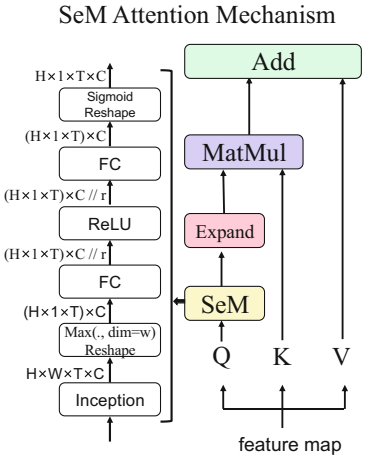


Fig. 5. The overview of SeM Attention Mechanism. The attention mechanism is widely used in Transformer.

3.4 Temporal Feature Aggregator

We propose an effective and low-consumption module, Temporal Feature Aggregator (TFA). The TFA module includes three layers, 'Conv2d_down-Conv2d_inter-Conv2d_up'.

Suppose the input to the module is $X_t \in \mathbb{R}^{C_{begin}\times T\times H\times W}$, where C_{begin} indicates the total number of channels of input, T represents the length of the input gait sequence and (H, W) denotes the height and width of each silhouette image. Therefore, the Temporal Feature Aggregator can be formulated as:

$$Y_{beg} = Max(X_t, dim = W) \tag{3}$$

$$Y_{mid} = F_r(F_b(F_u(F_{br}(F_i(F_{br}(F_d(Y_{beg})))))) + Y_{beg}) \tag{4}$$

$$Y_{end} = GMP(Y_{mid}) \tag{5}$$

where $Max(., dim = W)$ represents the maximization of the input along the wide(W) dimension, and $Y_{beg} \in \mathbb{R}^{C_{begin}\times T\times H\times 1}$ is the first output. $F_d(.)$, $F_i(.)$

and $F_u(.)$ denote three different 2d convolutions with output channel C_{down}, C_{inter} and C_{begin}. Their convolution kernel sizes are $(1,1)$, $(T_i,1)$ and $(1,1)$ respectively. $F_b(.)$ and $F_r(.)$ represent batch normalization and ReLU operations respectively, and $F_{br}(.)$ denotes that the input performs the batch normalization operation first and then the ReLU operation. $GMP(.)$ denotes global maximum pooling operation. $Y_{mid} \in \mathbb{R}^{C_{begin} \times T \times H \times 1}$ and $Y_{end} \in \mathbb{R}^{C_{begin} \times 1 \times 1 \times 1}$ denotes the output of the corresponding step.

3.5 Generalized-Mean Horizontal Pyramid Pooling

GeMHPP is an in-between method, determined by parameter learning. The GeMHPP module can be represented as:

$$Y_{GeMHPP} = (F_{Avg}(Y_{GeMin}.pow(p)))).pow(1/p) \tag{6}$$

where Y_{GeMin} and Y_{GeMHPP} denotes the input and output of GeMHPP. F_{Avg} denotes the average pooling operation. $pow(.)$ denotes the power operation. p is a parameter to be learned, and a suitable parameter is derived by multiple training.

In order to better train the proposed model, we use both triplet loss [8] and cross entropy loss.

4 Experiments

In this section, we first compare the experimental results on CASIA-B dataset with the state-of-the-art methods, then perform ablation study to compare the influence of different modules. Then we compare the experimental results on OUMVLP dataset.

4.1 Datasets

(1) CASIA-B [28] is a multi-view large gait dataset. There are gait samples of 124 subjects in this dataset. The samples are collected from 11 views (0°, 18°, …, 162°, 180°).
(2) OUMVLP [20] contains above 10,000 subjects. Each subject's samples are collected from 14 views (0°, 15°, 30°, 45°, 60°, 75°, 90°, 180°, 195°, 210°, 215°, 240°, 255°, and 270°).
(3) GREW [33] is currently recognized as the most extensive gait dataset in the wild according to available information. It consists of raw videos collected from 882 cameras positioned in a large public area, resulting in a substantial collection of nearly 3,500 h of video streams at a resolution of $1,080 \times 1,920$.

Table 1. Rank-1 accuracy (%) on CASIA-B dataset under all view angles, different settings and conditions

Gallery NM#1-4 Probe		Gallery view: 0° – 180°											Mean
		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	
NM#5-6	ACL [30]	92	98.5	**100**	**98.9**	95.7	91.5	94.5	97.7	98.4	96.7	91.9	96
	GEINet [19]	40.2	38.9	42.9	45.6	51.2	42	53.5	57.6	57.8	51.8	47.7	48.1
	GaitSet [2]	90.8	97.9	99.4	96.9	93.6	91.7	95	97.8	98.9	96.8	85.8	95
	GaitPart [7]	94.1	98.6	99.3	98.5	94	92.3	95.9	98.4	99.2	97.8	90.4	96.2
	GLFE [14]	96	98.3	99	97.9	**96.9**	95.4	97	98.9	99.3	98.8	94	97.4
	GLN [9]	93.2	**99.3**	99.5	98.7	96.1	95.6	97.2	98.1	99.3	98.6	90.1	96.9
	Ours	**96.1**	98.5	99.1	97.9	96.4	**95.6**	**97.4**	**98.9**	**99.3**	**99**	**95.5**	**97.6**
BG#1-2	GEINet [19]	34.2	29.3	31.2	35.2	35.2	27.6	35.9	43.5	45	39	36.8	35.7
	GaitSet [2]	83.8	91.2	91.8	88.8	83.3	81	84.1	90	92.2	94.4	79.0	87.2
	GaitPart [7]	89.1	94.8	96.7	95.1	88.3	84.9	89	93.5	96.1	93.8	85.8	91.5
	GLFE [14]	92.6	96.6	96.8	95.5	93.5	89.3	92.2	96.5	**98.2**	96.9	91.5	94.5
	GLN [9]	91.1	**97.7**	**97.8**	95.2	92.5	**91.2**	**92.4**	96	97.5	95	88.1	94
	Ours	**92.7**	95.5	97	**95.6**	**94.3**	88.8	92.2	**96.8**	97.9	**97.2**	**92.2**	**94.6**
CL#1-2	GEINet [19]	19.9	20.3	22.5	23.5	26.7	21.3	27.4	28.2	24.2	22.5	21.6	23.45
	GaitSet [2]	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50	70.4
	GaitPart [7]	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	GLFE [14]	76.6	90	90.3	87.1	84.5	**79**	84.1	87	87.3	84.4	69.5	83.6
	GLN [9]	70.6	82.4	85.2	82.7	79.2	76.4	76.2	78.9	77.9	78.7	64.3	77.5
	Ours	**78.7**	**90.8**	**92.7**	**90.2**	**85.1**	78.9	**84.3**	**87.5**	**89.1**	**86.7**	**72.9**	**85.2**

4.2 Experiment Results on CASIA-B

To test the performance in cross-view scenarios, we compare GaitSE with the latest advanced methods. As shown in Table 1, GaitSE outperforms SOTA in most views. Specifically, GaitSE outperforms previous methods by at least 0.2%, 0.1% and 1.6% in three conditions (normal/bag/cloth). Most notably, in the most challenging CL condition, GaitSE achieves an accuracy of 85.2%, a 1.6% improvement compared to GaitGL [14], which validates the robustness of GaitSE in difficult scenarios.

From the results we see that the accuracy of our proposed network can be greatly improved under cl conditions, which proves that the potential of the model can be improved by SeM, thus improving the resistance of the network to interference. From the graphs, we can see that the accuracy improvement is much larger in the (0°, 180°) than in the other angles.

4.3 Experiment Results on OUMVLP

In this section, we evaluate the performance of our proposed model on a larger OUMVLP dataset. The experimental settings in this section follow the setup of GaitPart [7] and GaitGL [14]. In Table 2, we evaluate gait samples at 14 different

Table 2. The recognition accuracy (%) comparison on OUMVLP dataset under 14 probe views excluding identical views.

Method	Gallery view: 0° – 180°														
	0°	15°	30°	45°	60°	75°	90°	108°	195°	210°	225°	240°	255°	270°	Mean
GEINet [19]	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet [2]	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1
GaitPart [7]	82.6	88.9	90.8	91.0	89.7	89.9	89.5	85.2	88.1	90.0	90.1	89.0	89.1	88.2	88.7
GLN [9]	83.8	90.0	91.0	91.2	90.3	90.0	89.4	85.3	89.1	90.5	90.6	89.6	89.3	88.5	89.2
GLFE [14]	84.9	90.2	91.1	91.5	91.1	90.8	90.3	88.5	88.6	90.3	90.4	89.6	89.5	88.8	89.7
Ours	**86.6**	**90.8**	**91.4**	**91.7**	**91.5**	**91.1**	**90.7**	**89.8**	**89.3**	**90.6**	**90.7**	**90.2**	**90**	**89.5**	**90.3**

views. During the test, we used Seq#00 and Seq#01 as the probe and gallery sequence, respectively. From the results, it seems that our proposed method improves more in the larger scale dataset than in the smaller scale dataset.

4.4 Experiment Results on GREW

We have analyzed the effectiveness of the proposed method by comparing its performance with various gait recognition methods using the GREW dataset. The evaluated methods, namely GaitGraph [21], GaitSet [2], Gaitpart [7], GaitGL [14], and CSTL [11], have been thoroughly assessed, and their respective experimental outcomes are presented in Table 3. Our findings from this comparison reveal an important trend. It appears that the gait recognition methods, which demonstrate satisfactory results in controlled laboratory settings, exhibit a notable decline in performance when confronted with real-world scenarios and datasets.

Table 3. The recognition accuracy (%) comparison on GREW dataset

Method	R-1 %	R-5 %	R-10 %
GaitGraph [21]	6.25	16.23	5.18
GaitSet [2]	46.3	63.6	70.3
GaitPart [7]	44	60.7	67.3
GaitGL [14]	47.3	63.6	69.3
CSTL [11]	**50.6**	65.9	71.9
ours	49	**66.2**	**72.4**

4.5 Training Details

The alignment of the input silhouettes is referred to [20], and the resolution size of the final silhouette is 64×44. Adam as our optimizer sets the learning rate and momentum to 1e−4 and 0.9, respectively. The margin m in the Eq. ?? about triplet loss is set to 0.2. The length of the gait sequences T is set to 30. Four NVIDIA 3080TI GPUs are used as our computational resources to train our model.

4.6 Ablation Study

We design various pertinent ablation experiments to analyze the importance of different modules.

Analysis of Global and Local Feature Extractor Based on SeNet. In GLSE, the role of the SEM module is mainly to reorganize features. In GaitGL [14], only local features are fused together to form a new gait feature map ignoring the connections between regions, the SEM module helps to establish the interactions between regions (Table 4).

Table 4. The recognition accuracy (%) of different max-pooling strategies in SE module on CASIA-B dataset.

SEM			NM	BG	CL	MEAN
Height	Width	Temporal				
✓	×	×	97	94.1	83.6	91.6
×	✓	×	**97.6**	**94.6**	**85.2**	**92.5**
✓	✓	×	97.3	94.1	83.7	91.7
✓	✓	✓	97.4	94.4	84.5	92.1

Analysis of Temporal Feature Aggregator. We set the TFA in different places and set different temporal feature convolution kernel sizes, and drew experimental conclusions in Table 5. In order to fully verify the best hyperparameters, we set two positions, which are after GLSEA2 and GLSEC. Meanwhile, we also set three convolution kernel sizes, i.e., (3, 1), (4, 1) and (5, 1). From the table we can see that the hyperparameters mentioned above have quite a strong influence on the model, especially under the CL condition.

Table 5. The recognition accuracy (%) of placing TFA in different positions on CASIA-B dataset.

TFA		NM	BG	CL	MEAN
Location	kernel size				
GLSEA2	(3, 1)	97.3	94.2	84	91.8
GLSEA2	(4, 1)	97.3	94.6	84.1	92
GLSEA2	(5, 1)	97.1	94	83.8	91.6
GLSEC	(3, 1)	97.4	94.5	84.2	92
GLSEC	(5, 1)	97.2	94.3	84.1	91.9
GLSEC	(4, 1)	**97.6**	**94.6**	**85.2**	**92.5**

5 Conclusions

In this paper, we propose a novel gait recognition framework that is capable of enhancing temporal global and local gait information, which can better generate interactions among the local regions and thus improve the robustness of the gait recognition task. First, we propose to partition the features into multiple local regions along the temporal dimension to extract discriminative features separately, i.e., Temporal Global and Local Aggregator. Second, we propose to introduce SEM into the local features in order to better utilize the local features, which enhances the interaction between regions. Our experiments on public datasets including both CASIA-B and OUMVLP demonstrate the superiority of our proposed framework.

Acknowledgment. This work was supported by National Natural Science Foundation of China (Grant Nos. 61906074, 32371984), Guangdong Basic and Applied Basic Research Foundation (Grant No. 2019A1515011276), Guangzhou Basic and Applied Basic Research Foundation (Grant No. 2023A04J1669), Key-Area Research and Development Program of Guangdong Province (Grant No. 2019B020214003, 2023B0202090001), China Agriculture Research System (CARS-15-23).

References

1. Ben, X., Gong, C., Zhang, P., Yan, R., Wu, Q., Meng, W.: Coupled bilinear discriminant projection for cross-view gait recognition. IEEE Trans. Circuits Syst. Video Technol. **30**(3), 734–747 (2019)
2. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: regarding gait as a set for cross-view gait recognition. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 8126–8133 (2019)
3. Chao, H., Wang, K., He, Y., Zhang, J., Feng, J.: Gaitset: cross-view gait recognition through utilizing gait as a deep set. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
4. Chen, X., Luo, X., Weng, J., Luo, W., Li, H., Tian, Q.: Multi-view gait image generation for cross-view gait recognition. IEEE Trans. Image Process. **30**, 3041–3055 (2021)
5. Chen, X., Weng, J., Lu, W., Xu, J.: Multi-gait recognition based on attribute discovery. IEEE Trans. Pattern Anal. Mach. Intell. **40**(7), 1697–1710 (2018)
6. Chen, X., Xu, J.: Uncooperative gait recognition: re-ranking based on sparse coding and multi-view hypergraph learning. Pattern Recogn. **53**, 116–129 (2016)
7. Fan, C., et al.: GaitPart: temporal part-based model for gait recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14225–14233 (2020)
8. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
9. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: learning discriminative and compact representations for gait recognition. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12354, pp. 382–398. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58545-7_22
10. Huang, T., Ben, X., Gong, C., Zhang, B., Yan, R., Wu, Q.: Enhanced spatial-temporal salience for cross-view gait recognition. IEEE Trans. Circuits Syst. Video Technol. **32**(10), 6967–6980 (2022)

11. Huang, X., et al.: Context-sensitive temporal feature learning for gait recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 12909–12918 (2021)
12. Li, G., Guo, L., Zhang, R., Qian, J., Gao, S.: Transgait: multimodal-based gait recognition with set transformer. Appl. Intell. **53**(2), 1535–1547 (2023)
13. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: Proceedings of the Asian Conference on Computer Vision (2020)
14. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: IEEE International Conference on Computer Vision (ICCV), pp. 14648–14656 (2021)
15. Lin, B., Zhang, S., Yu, X., Chu, Z., Zhang, H.: Learning effective representations from global and local features for cross-view gait recognition. arXiv preprint arXiv:2011.01461, **1**(2) (2020)
16. Liu, W., Zhang, C., Ma, H., Li, S.: Learning efficient spatial-temporal gait features with deep learning for human identification. Neuroinformatics **16**(3), 457–471 (2018)
17. Qin, H., Chen, Z., Guo, Q., Wu, Q.J., Lu, M.: RPnet: gait recognition with relationships between each body-parts. IEEE Trans. Circuits Syst. Video Technol. **32**(5), 2990–3000 (2021)
18. Sepas-Moghaddam, A., Etemad, A.: View-invariant gait recognition with attentive recurrent learning of partial representations. IEEE Trans. Biometrics Behav. Identity Sci. **3**(1), 124–137 (2020)
19. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: view-invariant gait recognition using a convolutional neural network. In: International Conference on Biometrics (ICB) (2016)
20. Takemura, N., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. IPSJ Trans. Comput. Vision Appl. **10**(1), 4 (2018)
21. Teepe, T., Khan, A., Gilg, J., Herzog, F., Hörmann, S., Rigoll, G.: Gaitgraph: graph convolutional network for skeleton-based gait recognition. In: 2021 IEEE International Conference on Image Processing (ICIP), pp. 2314–2318. IEEE (2021)
22. Thapar, D., Nigam, A., Aggarwal, D., Agarwal, P.: VGR-net: a view invariant gait recognition network. In: 2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA), pp. 1–8. IEEE (2018)
23. Wolf, T., Babaee, M., Rigoll, G.: Multi-view gait recognition using 3D convolutional neural networks. In: 2016 IEEE International Conference on Image Processing (ICIP), pp. 4165–4169. IEEE (2016)
24. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep CNNs. IEEE Trans. Pattern Anal. Mach. Intell. **39**(2), 209–226 (2016)
25. Xing, W., Li, Y., Zhang, S.: View-invariant gait recognition method by three-dimensional convolutional neural network. J. Electron. Imaging **27**(1), 013010 (2018)
26. Xu, C., Makihara, Y., Li, X., Yagi, Y., Lu, J.: Cross-view gait recognition using pairwise spatial transformer networks. IEEE Trans. Circuits Syst. Video Technol. **31**(1), 260–274 (2020)
27. Xu, Z., Lu, W., Zhang, Q., Yeung, Y., Chen, X.: Gait recognition based on capsule network. J. Vis. Commun. Image Represent. **59**, 159–167 (2019)

28. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: 18th International Conference on Pattern Recognition (2006)
29. Zhang, K., Luo, W., Ma, L., Liu, W., Li, H.: Learning joint gait representation via quintuplet loss minimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4700–4709 (2019)
30. Zhang, Y., Huang, Y., Yu, S., Wang, L.: Cross-view gait recognition by discriminative feature learning. IEEE Trans. Image Process. **29**, 1001–1015 (2020)
31. Zhang, Z., Tran, L., Liu, F., Liu, X.: On learning disentangled representations for gait recognition. IEEE Trans. Pattern Anal. Mach. Intell. (2020)
32. Zhang, Z., et al.: Gait recognition via disentangled representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4710–4719 (2019)
33. Zhu, Z., et al.: Gait recognition in the wild: a benchmark. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 14789–14799 (2021)