GaitMask: Mask-based Model for Gait Recognition

Beibei Lin†
18126289@bjtu.edu.cn
Yu Liu†
18126294@bjtu.edu.cn
Shunli Zhang*
slzhang@bjtu.edu.cn

School of Software Engineering
Beijing Jiaotong University
Beijing, China

Abstract

Gait recognition is an important biometric technology that identifies a person by using walking posture. Recently, most gait recognition methods either take the human gait as a whole to generate Global Feature Representation (GFR) or equivalently divide the human gait into multiple local regions to establish Local Feature Representation (LFR). However, we observe that LFR or GFR does not adequately represent the human gait because that LFR only focuses on the detailed information of each local region and GFR pays more attention to the global context information. On the other hand, the partition manner of the local regions is fixed, which only focuses on the local information of several specific regions. Motivated by this observation, we propose a novel mask-based network, named GaitMask, for gait recognition. GaitMask is built based on the Mask-based Local Augmentation (MLA), which is used to learn more comprehensive feature representations. MLA is a dual-branch structure consisting of a GFR extraction module as the trunk and a mask-based LFR extraction module as the branch. Specifically, the mask-based LFR extraction consists of a pair of complementary masks, where one mask randomly drops a region of the input feature maps and the other one only preserves this region. The complementary mask can be used to generate more comprehensive LFR and enhances the robustness of feature representations of the trunk. Experiments on two popular datasets demonstrate that our method achieves state-of-the-art results. Specifically, the proposed method significantly increases the performance in complex environments.

1 Introduction

Different from traditional biometric technologies such as face, fingerprint and iris, gait recognition can be used in long-distance condition and does not need the cooperation of the subjects. Hence, it is widely applied in surveillance systems and identity authentication. However, the performance of gait recognition suffers from many complicated factors, including view, carrying and speed, etc [4, 23, 24]. Thus, gait recognition is still a challenging task.

†Joint first authors.
*Shunli Zhang is the corresponding author.

Recently, researchers propose different deep-learning based methods to generate discriminative feature representations, which can be roughly divided into two categories. One is Global Feature Representation (GFR) that treats human gait as a whole for feature extraction [3, 8, 10, 11, 12, 16, 17, 21, 22]. The other one is Local Feature Representation (LFR) that extracts gait features from multiple local regions [5, 9, 14, 15, 27].

However, we observe that LFR neglects the correlation of different local regions, while GFR does not take full advantage of detailed information. Moreover, as shown in Fig.1(b), the local regions are partitioned by a top-to-bottom, equal-size manner [5, 27], which only focuses on the information of a few specific localized regions.

To solve these problems mentioned above, in this paper, we propose a novel mask-based LFR extractor to generate more comprehensive LFR. Specifically, the mask-based LFR extractor is a dual-branch structure including a pair of complementary masks. As shown in fig.1(c), one mask randomly drops a local region of gait sequences, while the other one only preserves this region. During the training stage, by using the complementary masks, this extractor randomly generates a pair of complementary feature maps that can be used to extract local gait features from arbitrary local regions. Compared to other local partition patterns that can only utilize information from a few specific regions, the proposed extractor takes full advantage of the information from different local regions to train the network. Hence, this extractor can generate more comprehensive LFR during the test stage. Based on the mask-based LFR extractor, we propose a novel feature extraction module, called Mask-based Local Augmentation (MLA), to generate more comprehensive feature representations. MLA includes a GFR extractor and a mask-based LFR extractor. GFR extractor is used as the trunk to generate GFR from the whole feature maps, while mask-based LFR extractor generates LFR to enhance the feature representation of the trunk.

The main contributions of our method can be summarized as follows:

- We propose a novel LFR extractor, which can be used to generate more comprehensive LFR by using a pair of complementary masks. Unlike traditional partitions that only extract features from several fixed local regions, the proposed LFR extractor efficiently utilizes gait information from different local regions.

- Based on the proposed mask-based LFR extractor, we develop a novel Mask-based Local Augmentation, consisting of a trunk and a branch, to generate more discriminative feature representations. The trunk focuses on global context information, while the branch pays more attention to the detailed information of gait sequences.

- The experimental results on two benchmark datasets demonstrate the proposed method achieves State-Of-The-Art (SOTA) results. Specifically, our method outperforms other methods by 1.7% and 5.5% in the carrying condition of bag and coat, respectively.

2 Related Work

Recently, most deep-learning based gait recognition methods take the silhouettes of gait sequences as input to extract gait features by using 2D or 3D convolutional neural networks (CNNs). These methods can be roughly divided into two types, *i.e.*, template-based and sequence-based.

The template-based methods either aggregate temporal information of gait sequences as Gait Energy Image (GEI) to extract gait features or extract each gait image's features and

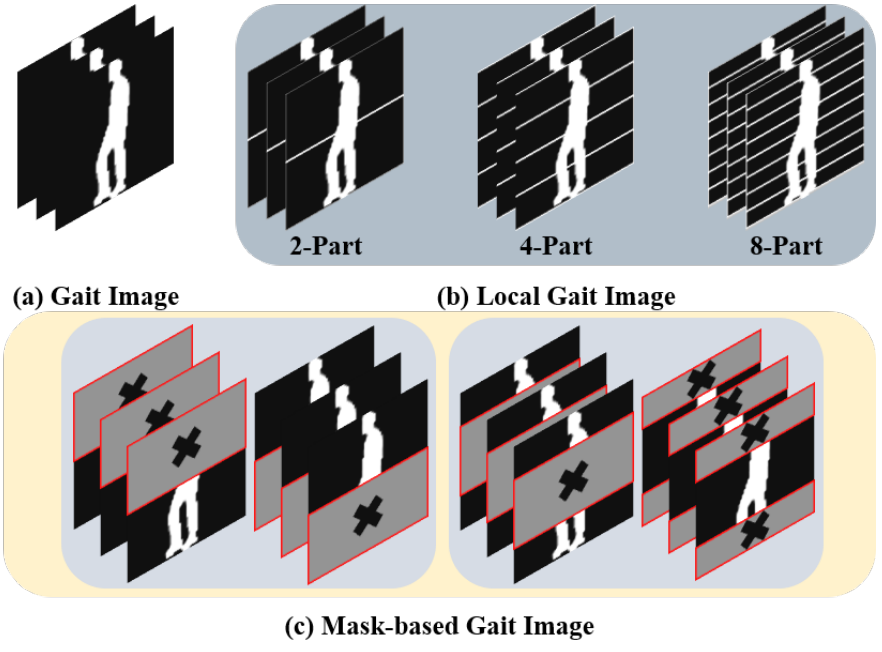

Figure 1: Visualization of the original gait image, local gait images of different partitions and mask-based gait image.

then integrate temporal information [1, 3, 5, 7, 16, 22, 25]. For example, Shiraga et al. [16] propose a template-based network named GEINet to generate feature representations from GEI. Specifically, they first generate GEI by using the mean function to aggregate all temporal information of gait sequences and then extract gait features with 2D CNN. However, the generation process of GEI leads to the loss of a large amount of information. To better leverage the information of gait sequences, some researchers [3, 5, 27] first use 2D CNN to extract gait features of each gait image and then aggregate temporal information of gait sequences. For example, Chao et al.[3] propose a novel network named GaitSet to generate discriminative feature representations. Gaitset first extracts each gait image's gait features and then uses the max function to aggregate temporal information. However, they cannot fully utilize the temporal information of gait sequences. To better take advantage of this information, some researchers [5, 27] model temporal relationship after spatial feature extraction. For example, Fan et al.[5] propose a Micro-Motion Capture Module (MCM) to model the short-range temporal dependence. Zhang et al. [27] use the long short-term memory (LSTM) units to model temporal relationships.

The sequence-based methods usually take gait sequences as a unit to extract spatial-temporal gait features by using 3D CNN [12, 13, 19, 20]. For example, Wolf et al.[20] develop a 3D CNN to generate spatial-temporal gait representations from a fixed-length gait clip. Thapar et al. [19] first partition gait sequences into multiple fixed-length gait clips and then use 3D CNN to extract gait features of each clip. Finally, they use an LSTM module to learn the temporal relationship of different gait clips. However, these works are inflexible because they need a fixed-length gait clip as input to train their network. To make full use of the temporal information in 3D CNNs, Lin et al.[12] propose a frame pooling operation to

aggregate adaptively the temporal information of the whole gait sequences, which take full advantage of temporal information of a whole gait sequence.

3 Proposed Method

In this section, we first outline the framework of our GaitMask method. Then, we introduce the proposed Mask-based Local Augmentation (MLA). Finally, we present the implementation details of the training and test phases.

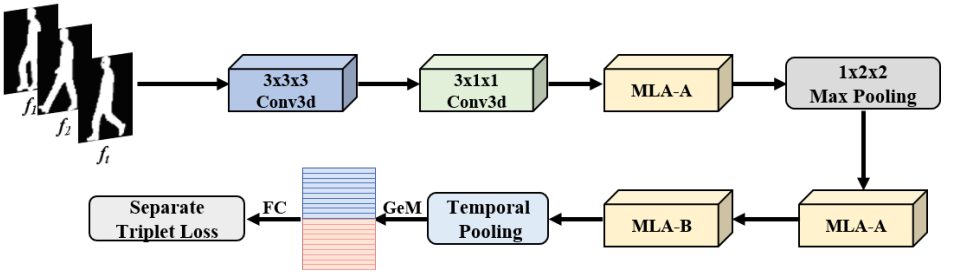

Figure 2: Overview of the proposed GaitMask.

3.1 Overview

The overview of our GaitMask method is shown in Fig. 2. The whole gait recognition method is built by 3D convolution, which is similar to [12]. Given a gait sequence, we first use a 3D convolution to extract the shallow features. Then, a temporal convolution is used to aggregate the local temporal information of feature maps. Next, multiple MLA modules are proposed to learn more comprehensive gait features. Finally, we introduce temporal pooling and Generalized Mean Pooling (GeM) to generate feature representations. During the training stage, we use the separate triplet loss to train the proposed network [4, 5].

3.2 Mask-based Local Augmentation

As shown in Fig.3, MLA includes two branches: GFR extraction and Mask-based LFR extraction. GFR extraction extracts gait features directly from the input feature map, while Mask-based LFR extraction first generates a pair of complementary gait feature maps and then extracts local gait features from them. Assume that the input feature map of MLA is $X_{in} \in \mathbb{R}^{C_{in} \times T_{in} \times H_{in} \times W_{in}}$, where C_{in} is the number of channels, T_{in} is the length of feature maps and (H_{in},W_{in}) is the image size of each frame. GFR extraction can be defined as

$$Y_g = c^{3\times3\times3}(X_{in}), \tag{1}$$

where $c^{3\times3\times3}$ means 3D convolution with kernel size 3. $Y_g \in \mathbb{R}^{C_{ou} \times T_{in} \times H_{in} \times W_{in}}$ is the output of GFR extraction.

On the other hand, Mask-based LFR extraction first generates two complementary masks $M_0 \in \mathbb{R}^{H_{in} \times W_{in}}$ and $M_1 \in \mathbb{R}^{H_{in} \times W_{in}}$, where the element of M_0 and M_1 is 0 and 1, respectively. Then, we randomly drop a continuous and horizontal region of the mask M_1. Meanwhile, we preserve the corresponding region in the mask M_0. Specifically, assume that

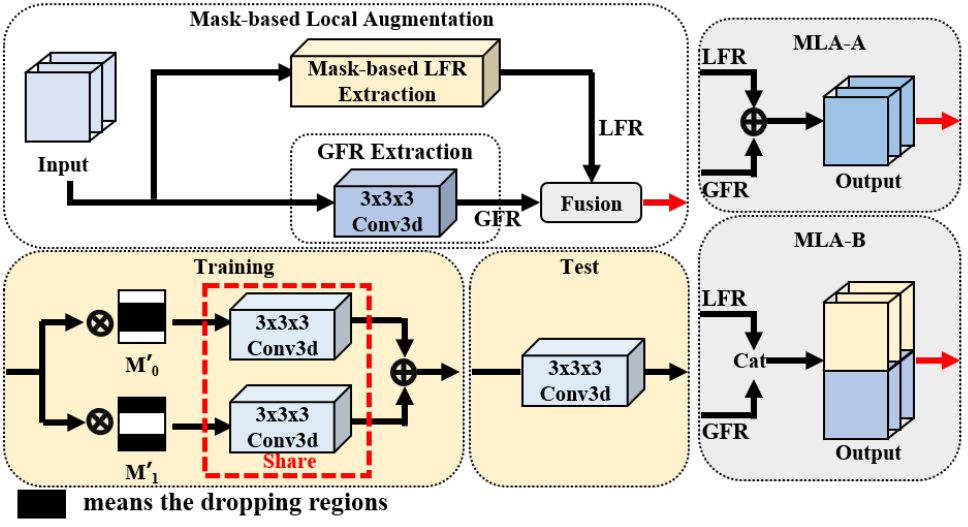

Figure 3: Overview of the proposed Mask-based Local Augmentation. Mask-based Local Augmentation includes two branches: GFR Extraction and Mask-based LFR Extraction. After GFR and LFR Extractions, we propose two different operations to fuse both feature representations, called MLA-A and MLA-B. MLA-A means element-wise addition, while MLA-B means concatenating two feature maps horizontally. During the training stage, the input of the Mask-based LFR Extraction is a pair of complementary mask-based feature maps. During the test stage, the input of the Mask-based LFR Extraction is the original feature maps.

$M_0 = \{h_i^0 | i = 1,2,...,H_{in}\}$, where $h_i^0 \in \mathbb{R}^{1\times W_{in}}$ is the i-th column of the feature map M_0. $M_1 = \{h_j^1 | j = 1,2,...,H_{in}\}$, where $h_j^1 \in \mathbb{R}^{1\times W_{in}}$ is the j-th column of the feature map M_1. We first randomly select an interval $(k, k+\lfloor d \times H_{in} \rfloor)$, where d means the dropping ratio. Then, the value of $\{h_k^0,...,h_{k+\lfloor d\times H_{in}\rfloor}^0\}$ in the mask M_0 is set to 1, as a new mask $M_0^{'} \in \mathbb{R}^{H_{in}\times W_{in}}$, while the value of $\{h_k^1,...,h_{k+\lfloor d\times H_{in}\rfloor}^1\}$ is set to 0, as a new mask $M_1^{'} \in \mathbb{R}^{H_{in}\times W_{in}}$. Mask-based LFR extraction can be represented as:

$$Y_m = c^{3\times3\times3}(\sum_{k=1}^{C_{in}}\sum_{q=1}^{T_{in}}(X_{in}\otimes M_0^{'})) + c^{3\times3\times3}(\sum_{k=1}^{C_{in}}\sum_{q=1}^{T_{in}}(X_{in}\otimes M_1^{'})), \tag{2}$$

where $\otimes$ means element-wise product in the image dimension. $Y_m \in \mathbb{R}^{C_{ou}\times T_{in}\times H_{in}\times W_{in}}$ is the output of mask-based LFR extraction.

In this paper, we propose two ways to combine the outputs of two extractions. One is the element-wise addition (MLA-A), which can be designed as

$$Y_{MLA-A} = Y_g + Y_m, \tag{3}$$

where $Y_{MLA-A} \in \mathbb{R}^{C_{ou}\times T_{in}\times H_{in}\times W_{in}}$ is the combined feature maps. The other one is to concatenate the feature maps in horizontal axis, which can be represented as

$$Y_{MLA-B} = concat\begin{Bmatrix} Y_g \\ Y_m \end{Bmatrix}, \tag{4}$$

where *concat* means concatenation operation in horizontal axis. $Y_{MLA-B} \in \mathbb{R}^{C_{ou} \times T_{in} \times (2*H_{in}) \times W_{in}}$ is the combined feature maps.

3.3 Feature Mapping

After feature extraction, we introduce temporal pooling and spatial pooling to generate feature representations. The temporal pooling aims to aggregate all temporal information of gait sequences [3, 5, 12]. Assume that $X_{fm} \in \mathbb{R}^{C_{fin} \times T_{fin} \times H_{fin} \times W_{fin}}$ is the output of the last MLA module, where C_{fin} is the number of channels, T_{fin} is the length of feature maps and (H_{fin},W_{fin}) is the image size of each frame. The temporal pooling can be defined as

$$Y_{tp} = F_{Max}^{T_{fin} \times 1 \times 1}(X_{fm}), \tag{5}$$

where $F_{Max}^{T_{fin} \times 1 \times 1}(\cdot)$ means max-pooling operation. $Y_{tp} \in \mathbb{R}^{C_{fin} \times 1 \times H_{fin} \times W_{fin}}$ is the output of temporal pooling.

The spatial pooling first partitions the feature map Y_{tp} into multiple horizontal strips and then uses the Generalized-Mean pooling (GeM) to aggregate adaptively each strip's information in the vertical axis. Finally, multiple separate fully connected layers are used to further integrate the channel information of each strip [3, 5]. The spatial pooling can be represented as

$$Y_{sp} = F_{sfc}((F_{Avg}^{1 \times 1 \times W_{fin}}((Y_{tp})^p))^{\frac{1}{p}}), \tag{6}$$

where $F_{Avg}^{1 \times 1 \times W_{fin}}(\cdot)$ means average-pooling operation. F_{sfc} means multiple separate Fully Connected (FC) layers. Its size is $H_{fin} \times C_{fin} \times C_{fou}$, where H_{fin} is the number of FC layers, and C_{fin} and C_{fou} are the input and output dimensions of each FC layer, respectively. $Y_{tp} \in \mathbb{R}^{C_{fou} \times 1 \times H_{fin} \times 1}$ is the output of spatial pooling.

3.4 Training Details and Test

Training. During the training phase, we first randomly crop a gait clip as an input of the network. Then, the feature representation Y_{tp} will be generated. Finally, the separate triplet loss is used to calculate the loss of each strip independently [3, 5, 12]. The triplet loss can be defined as:

$$L_{triplet} = Max(D(Y_{tp}^{\alpha}, Y_{tp}^{\beta}) - D(Y_{tp}^{\alpha}, Y_{tp}^{\gamma}) + margin, 0) \tag{7}$$

where α and β are samples from the same class, while γ represents samples from another class. $D_{(d_i,d_j)}$ is the Euclidean Distance between the sample d_i and d_j. *margin* is the margin of the triplet loss. To better train the proposed network, we take the Batch ALL (BA) strategy as the sampling strategy[3]. Specifically, the number of samples per batch is set to $P \times K$, where P is the number of subject IDs and K is the number of samples per subject ID. Due to the limitation of memory size and computational complexity, the length of the input gait clip is set to T frames.

Test. During the test phase, the network does not need to consider the limitation of memory size. Hereby, the whole gait sequence can be fed into the proposed GaitMask to generate the feature representation $Y_{tp} \in \mathbb{R}^{C_{fou} \times 1 \times H_{fin} \times 1}$. Then, we flatten the feature representation Y_{tp} into a feature vector with dimension $C_{fou} \times H_{fin}$. To evaluate the performance of our method, we adopt the gallery-probe mode to calculate Rank-1 accuracy [3, 12].

Datasets	Training Set	Test Set	Gallery Set	Probe Set
CASIA-B	74	50	NM #01-04	NM#05-06 BG#01-02 CL#01-02
OUMVLP	5,153	5,154	Seq #01	Seq #00

Table 1: The evaluation protocol of CASIA-B and OUMVLP datasets.

4 Experiments

4.1 Datasets and Evaluation Protocol

CASIA-B. The CASIA-B dataset [24] is one of the most complex gait datasets, which includes 124 subjects. Each subject was collected in 10 groups of gait sequences (Normal walking (NM) #01-06, Walking with a bag (BG) #01-02 and Walking with a coat (CL) #01-02). Each group of gait sequences includes 11 view angles (0°,18°,..., 180°). In this paper, we use the same protocol as [5] to evaluate the performance of our method on CASIA-B dataset, which is shown in Tab.1.

OUMVLP. The OUMVLP dataset [18] is one of the largest gait datasets. It includes 10,307 subjects, each of which contains 2 groups of gait sequences (Seq#00-01). Each group includes 14 view angles (0°,15°,...,90° and 180°,195°,...,270°). In this paper, we adopt the same protocol as [2, 3] for evaluation, which is shown in Tab.1.

CASIA-B				OUMVLP			
Layer Name	In_C	Out_C	Kernel	Layer Name	In_C	Out_C	Kernel
Conv3d	1	32	(3, 3, 3)	Conv3d	1	64	(3, 3, 3)
				Conv3d	64	64	(3, 3, 3)
Conv3d	32	32	(3, 1, 1)	Conv3d	64	64	(3, 1, 1)
MLA-A	32	32	(3, 3, 3)	MLA-A	64	128	(3, 3, 3)
				MLA-A	128	128	(3, 3, 3)
Max Pooling	-	-	(1, 2, 2)	Max Pooling	-	-	(1, 2, 2)
MLA-A	64	128	(3, 3, 3)	MLA-A	128	196	(3, 3, 3)
				MLA-A	196	196	(3, 3, 3)
MLA-B	128	128	(3, 3, 3)	MLA-A	196	256	(3, 3, 3)
				MLA-B	256	256	(3, 3, 3)

Table 2: Network parameters of the proposed method on two datasets. In_C, Out_C and Kernel mean the input channels, output channels and kernel size, respectively.

4.2 Implementation Details

For all experiments of both datasets, we preprocessed and aligned the gait image into the same size 64×44 [3]. The margin in Equ.7 is set to 0.2. Adam is treated as the optimizer and the initialized learning rate is set to 1e-4. During the training phase, the length of the input sequence T is set to 30. The network parameters are shown in Tab.2. For the CASIA-B dataset, the batch size $P \times K$ is set to 8×16. The iteration is set to 80k and the learning rate is reset to 1e-5 for the last 10K iterations. For the OUMVLP dataset, the parameters P and K are set as 32 and 16, respectively. The iteration is set as 250K and the learning rate is reset to 1e-5 after 150K iterations.

Gallery NM#1-4		0°-180°											
Probe		0°	18°	36°	54°	72°	90°	108°	126°	144°	162°	180°	mean
NM#5-6	CNN-3D	87.1	93.2	97.0	94.6	90.2	88.3	91.1	93.8	96.5	96.0	85.7	92.1
	CNN-Ensemble	88.7	95.1	98.2	96.4	94.1	91.5	93.9	97.5	98.4	95.8	85.6	94.1
	GaitSet	90.8	97.9	99.4	96.9	93.6	91.7	95.0	97.8	98.9	96.8	85.8	95.0
	ACL	92.0	98.5	**100.0**	**98.9**	95.7	91.5	94.5	97.7	98.4	96.7	91.9	96.0
	GaitPart	94.1	**98.6**	99.3	98.5	94.0	92.3	95.9	98.4	**99.2**	97.8	90.4	96.2
	Ours	**94.8**	97.5	98.9	97.3	**96.2**	**95.3**	**97.1**	**98.7**	98.5	**98.2**	**92.0**	**96.8**
BG#1-2	CNN-LB	64.2	80.6	82.7	76.9	64.8	63.1	68.0	76.9	82.2	75.4	61.3	72.4
	GaitSet	83.8	91.2	91.8	88.8	83.3	81.0	84.1	90.0	92.2	94.4	79.0	87.2
	GaitPart	89.1	94.8	96.7	**95.1**	88.3	84.9	89.0	93.5	96.1	93.8	85.8	91.5
	Ours	**90.7**	**95.2**	**95.9**	94.1	**92.5**	**87.2**	**91.6**	**95.2**	**97.4**	**96.7**	**88.5**	**93.2**
CL#1-2	CNN-LB	37.7	57.2	66.6	61.1	55.2	54.6	55.2	59.1	58.9	48.8	39.4	54.0
	GaitSet	61.4	75.4	80.7	77.3	72.1	70.1	71.5	73.5	73.5	68.4	50.0	70.4
	GaitPart	70.7	85.5	86.9	83.3	77.1	72.5	76.9	82.2	83.8	80.2	66.5	78.7
	Ours	**79.4**	**90.7**	**92.4**	**87.6**	**83.1**	**79.1**	**84.2**	**86.7**	**88.0**	**84.0**	**71.0**	**84.2**

Table 3: Rank-1 accuracy (%) on CASIA-B under all view angles and different conditions, excluding identical-view case.

Method	Probe View														Mean
	0°	15°	30°	45°	60°	75°	90°	180°	195°	210°	225°	240°	255°	270°	
GEINet	23.2	38.1	48.0	51.8	47.5	48.1	43.8	27.3	37.9	46.8	49.9	45.9	45.7	41.0	42.5
GaitSet	79.3	87.9	90.0	90.1	88.0	88.7	87.7	81.8	86.5	89.0	89.2	87.2	87.6	86.2	87.1
GaitPart	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	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
GaitKMM	56.2	73.7	81.4	82.0	78.4	78.0	76.5	60.2	72.0	79.8	80.2	76.7	76.3	73.9	74.7
Ours	**86.7**	**90.8**	**91.3**	**91.6**	**91.4**	**91.1**	**90.8**	**89.9**	**89.2**	90.3	90.5	**90.0**	**89.9**	**89.4**	**90.2**
GEINet	24.9	40.7	51.6	55.1	49.8	51.1	46.4	29.2	40.7	50.5	53.3	48.4	48.6	43.5	45.3
GaitSet	84.5	93.3	96.7	96.6	93.5	95.3	94.2	87.0	92.5	96.0	96.0	93.0	94.3	92.7	93.3
GaitPart	-	-	-	-	-	-	-	-	-	-	-	-	-	-	95.1
GLN	89.3	95.8	97.9	97.8	96.0	96.7	96.1	90.7	95.3	**97.7**	**97.5**	95.7	96.2	95.3	95.6
Ours	**92.3**	**96.7**	**98.2**	**98.3**	**97.3**	**97.9**	**97.6**	**95.7**	**95.5**	97.5	97.5	**96.2**	**96.9**	**96.4**	**96.7**

Table 4: Rank-1 accuracy (%) on OUMVLP dataset under different view angles, excluding identical-view cases. The last five rows show the results excluding invalid probe sequences.

4.3 Comparison with State-of-the-Art

Evaluation on CASIA-B. To evaluate the performance of our method in some complex situations, we conduct experiments on the CASIA-B dataset. The experimental results are shown in Tab.3, which includes several gait recognition methods, such as GaitSet [3], CNN-LB, CNN-3D, CNN-Ensemble [22], ACL [27] and GaitPart [5]. Experimental results demonstrate that the proposed method achieves the most excellent performance in almost all view angles. In particular, we observe that the proposed method obtains significant performance improvements compared to other methods under some unfavorable conditions. For example, the accuracy of GaitPart is 96.2% in the NM condition, while in the BG and CL cases, its accuracy is only 91.5% and 78.7%. In contrast, the accuracy of the proposed method is 96.8%, 93.2% and 84.2% in the NM, BG and CL conditions, respectively, which outperform the GaitPart by 0.6%, 1.7% and 5.5%, respectively. On the other hand, we also observe that the proposed method achieves better performance in some particular view angles, such as 90° and 180°. For example, the average accuracy of GaitMask in the NM condition is 96.8%, which outperforms the GaitPart by 0.6%. For 90° and 180°, the NM accuracy of GaitMask is 95.3% and 92.0%, which outperform the GaitPart by 3.0% and 1.6%, respectively. The main reason for the performance improvement is that the proposed method generates more comprehensive feature representations by taking full advantage of the local detailed information and global context information.

Evaluation on OUMVLP. Although it contains several complex conditions, CASIAB dataset has only 124 subjects. To evaluate the performance of our method in a larger dataset,

we carry out the experiment on OUMVLP dataset. As shown in Tab.4, we compare our method with several gait recognition methods, including GEINet [16], GaitSet [3], GaitPart [5], GLN[7] and GaitKMM[26]. It can be observed that the proposed method achieves the highest performance in almost all view angles. Specifically, we also observe that the proposed method significantly improves the recognition accuracy of some specific view angles. For example, the accuracy of the proposed method in 0° and 180° is 92.3% and 95.7%, respectively, which outperforms the GLN by 3.0% and 5.0%.

MLA		NM	BG	CL	Mean
GFE	LFE				
✓		96.1	91.9	81.4	89.8
	✓	96.1	91.8	81.9	89.9
✓	✓	96.8	93.2	84.2	91.4

Table 5: Rank-1 accuracy (%) of different feature extractions

4.4 Ablation Study

In this paper, we propose the GaitMask network, which includes Mask-based Local Augmentation module. To verify the effectiveness of the proposed MLA, we carry out several ablation studies on CASIA-B dataset.

Analysis of MLA module. In Sec.3.2, we propose a novel MLA module, which includes GFR extraction and Mask-based LFR extraction. GFR extraction aims to extract global context information, while LFR extraction is used to extract detailed information of gait sequences. To explore the contribution of each extraction, we design two ablation studies, each of which uses only one feature extraction module. The experimental results are shown in Tab.5. It can be observed that each feature extraction contributes to the overall recognition accuracy.

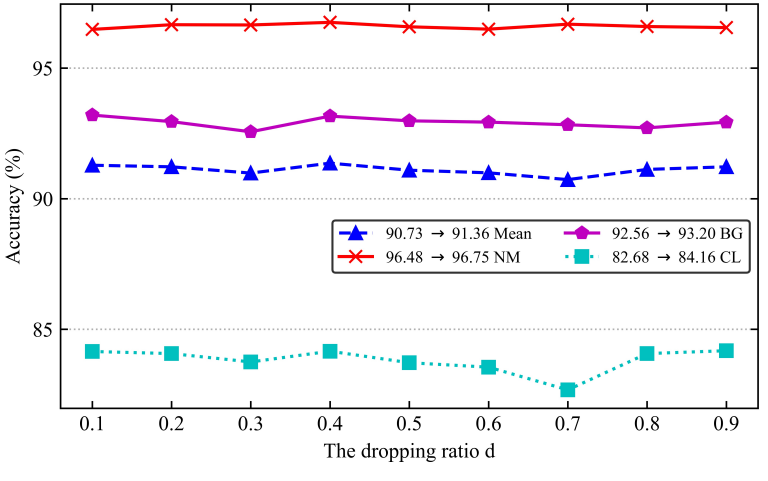

Figure 4: Rank-1 accuracy (%) of different dropping ratios.

Analysis of the dropping ratio d. The proposed MLA uses a pair of complementary masks to drop local regions of gait sequences, which can be used to extract more comprehensive LFR. In Sec.3.2, the dropping ratio is set to d. To explore the optimal value of d, we design

several experiments in which the range of values is $\{0.1, 0.2, ..., 0.9\}$. The experimental results are shown in Fig.4. It can be observed that the setting "d=0.4" achieves the highest recognition accuracy. Hereby, we set the value d to 0.4 during the training stage.

The number of the MLA	NM	BG	CL	Mean
0	96.1	91.9	81.4	89.8
1	96.4	92.3	83.3	90.7
2	96.2	92.7	83.4	90.8
3	96.8	93.2	84.2	91.4

Table 6: Rank-1 accuracy (%) of different MLA number

Analysis of the number of MLAs. To discuss the effect of the number of MLAs in the network, we design the ablation studies by using different numbers of MLAs. The experimental results are shown in Tab.6. It can be observed that larger number of MLAs can lead to higher recognition accuracy. Thereby, the number of MLAs on CASIA-B is finally set to three.

5 Conclusion

In this paper, we propose a novel perspective that utilizes a pair of complementary masks to extract more effective local feature representations from arbitrary local regions. Based on this perspective, we present a Mask-based Local Augmentation consisting of GFR and LFR extractions to generate more comprehensive gait feature representations. GFR extraction aims to capture the global context information, while LFR extraction is used to extract the detailed information of gait sequences. Experiments on two datasets demonstrate that the proposed method achieves state-of-the-art recognition accuracy.

Acknowledgements. This work was supported by the Beijing Natural Science Foundation (4202056), the National Natural Science Foundation of China (61976017 and 61601021), and the Fundamental Research Funds for the Central Universities (2020JBM078). The support and resources from the Center for High Performance Computing at Beijing Jiaotong University(http://hpc.bjtu.edu.cn) are gratefully acknowledged.

References

[1] Tianrui Chai, Xinyu Mei, Annan Li, and Yunhong Wang. Semantically-guided disentangled representation for robust gait recognition. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021.

[2] Tianrui Chai, Xinyu Mei, Annan Li, and Yunhong Wang. Silhouette-based view-embeddings for gait recognition under multiple views. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2319–2323. IEEE, 2021.

[3] Hanqing Chao, Yiwei He, Junping Zhang, and Jianfeng Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *AAAI*, volume 33, pages 8126–8133, 2019.

[4] Patrick Connor and Arun Ross. Biometric recognition by gait: A survey of modalities and features. *Computer Vision and Image Understanding*, 167:1–27, 2018.

[5] Chao Fan, Yunjie Peng, Chunshui Cao, Xu Liu, Saihui Hou, Jiannan Chi, Yongzhen Huang, Qing Li, and Zhiqiang He. Gaitpart: Temporal part-based model for gait recognition. In *CVPR*, pages 14225–14233, 2020.

[6] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[7] Saihui Hou, Chunshui Cao, Xu Liu, and Yongzhen Huang. Gait lateral network: Learning discriminative and compact representations for gait recognition. In *European Conference on Computer Vision*, pages 382–398. Springer, 2020.

[8] Saihui Hou, Xu Liu, Chunshui Cao, and Yongzhen Huang. Set residual network for silhouette-based gait recognition. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2021.

[9] Lily Lee and W Eric L Grimson. Gait analysis for recognition and classification. In *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*, pages 155–162. IEEE, 2002.

[10] Shuangqun Li, Wu Liu, and Huadong Ma. Attentive spatial–temporal summary networks for feature learning in irregular gait recognition. *IEEE Transactions on Multimedia*, 21(9):2361–2375, 2019.

[11] Xiang Li, Yasushi Makihara, Chi Xu, Yasushi Yagi, and Mingwu Ren. Gait recognition via semi-supervised disentangled representation learning to identity and covariate features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13309–13319, 2020.

[12] Beibei Lin, Shunli Zhang, and Feng Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *ACM MM*, pages 3054–3062, 2020.

[13] Beibei Lin, Shunli Zhang, Yu Liu, and Shengdi Qin. Multi-scale temporal information extractor for gait recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2998–3002. IEEE, 2021.

[14] Imad Rida, Xudong Jiang, and Gian Luca Marcialis. Human body part selection by group lasso of motion for model-free gait recognition. *IEEE Signal Processing Letters*, 23(1):154–158, 2015.

[15] Md Rokanujjaman, Md Altab Hossain, and Md Rezaul Islam. Effective part selection for part-based gait identification. In *2012 7th International Conference on Electrical and Computer Engineering*, pages 17–19. IEEE, 2012.

[16] Kohei Shiraga, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Geinet: View-invariant gait recognition using a convolutional neural network. In *ICB*, pages 1–8. IEEE, 2016.

[17] Chunfeng Song, Yongzhen Huang, Yan Huang, Ning Jia, and Liang Wang. Gaitnet: An end-to-end network for gait based human identification. *Pattern Recognition*, 96: 106988, 2019.

[18] Noriko Takemura, Yasushi Makihara, Daigo Muramatsu, Tomio Echigo, and Yasushi Yagi. Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition. *IPSJ Transactions on Computer Vision and Applications*, 10(1): 4, 2018.

[19] Daksh Thapar, Gaurav Jaswal, Aditya Nigam, and Chetan Arora. Gait metric learning siamese network exploiting dual of spatio-temporal 3d-cnn intra and lstm based inter gait-cycle-segment features. *Pattern Recognition Letters*, 125:646–653, 2019.

[20] Thomas Wolf, Mohammadreza Babaee, and Gerhard Rigoll. Multi-view gait recognition using 3d convolutional neural networks. In *ICIP*, pages 4165–4169. IEEE, 2016.

[21] Haoqian Wu, Jian Tian, Yongjian Fu, Bin Li, and Xi Li. Condition-aware comparison scheme for gait recognition. *IEEE Transactions on Image Processing*, 2020.

[22] Zifeng Wu, Yongzhen Huang, Liang Wang, Xiaogang Wang, and Tieniu Tan. A comprehensive study on cross-view gait based human identification with deep cnns. *IEEE TPAMI*, 39(2):209–226, 2016.

[23] TzeWei Yeoh, Hernán E Aguirre, and Kiyoshi Tanaka. Clothing-invariant gait recognition using convolutional neural network. In *ISPACS*, pages 1–5. IEEE, 2016.

[24] Shiqi Yu, Daoliang Tan, and Tieniu Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *ICPR*, volume 4, pages 441–444. IEEE, 2006.

[25] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu. Siamese neural network based gait recognition for human identification. In *ICASSP*, pages 2832–2836. IEEE, 2016.

[26] Shaoxiong Zhang, Yunhong Wang, and Annan Li. Cross-view gait recognition with deep universal linear embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9095–9104, June 2021.

[27] Yuqi Zhang, Yongzhen Huang, Shiqi Yu, and Liang Wang. Cross-view gait recognition by discriminative feature learning. *IEEE TIP*, 29:1001–1015, 2019.