Bag of Tricks and GaitMask Network for Gait Recognition

Beibei Lin[1] , Shunli Zhang[1]*, Jiande Sun[2], Shengdi Qin[1] and Chenwei Wan[1]
[1] Beijing Jiaotong University, [2] Shandong Normal University
{18126289, slzhang, 20121732, 20126339}@bjtu.edu.cn, jiandesun@hotmail.com

1. Method

The inference pipeline of our framework is shown in Fig. 1. Given a gait sequence, we first use the GaitGL [3] and GaitMask networks to generate gait feature representations. Then, we concatenate two feature representations. Finally, we utilize the query expansion and re-ranking [4] techniques to increase the accuracy.

1.1. GaitMask

The overview of the proposed GaitMask method is shown in Fig. 2. The whole gait recognition method is built by 3D convolution, which is similar to [2, 3]. We first use 3D convolution to extract the shallow features. Then, the Local Temporal Aggregation (LTA) is used to aggregate the local temporal information [3]. Next, we propose a novel Global and Mask Feature Extractor (GMFE) to learn more comprehensive gait features. Finally, we introduce temporal pooling and Generalized Mean Pooling (GeM) to generate feature representations [3]. During the training stage, we use the Separate Triplet Loss to train the proposed network [1].

The proposed GMFE can be implemented by multiple Global and Mask Convolution Layers (GMCL). The overview of the GMCL is shown in Fig.3. GMCL includes two branches: Global Feature Extraction and Mask Feature Extraction. Global Feature Extraction is used to extract global feature representations, while Mask Feature Extraction is used to generate more comprehensive local feature representations. Assume that the input feature map of the GMCL 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. The global feature extraction can be designed 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 the global feature extraction.

On the other hand, the mask feature 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 $M_0 = \left\{h_i^0 | i = 1, 2, ..., H_{in}\right\}$, where $h_i^0 \in \mathbb{R}^{1 \times W_{in}}$ is the *i*-th column of the feature map M_0. $M_1 = \left\{h_j^1 | j = 1, 2, ..., H_{in}\right\}$, 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 $(k, k + \frac{d * H_{in}}{H_{in}})$, where d means the dropping ratio. Then, the value of $\left\{h_k^0, ..., h_{k+\lfloor d \times H_{in} \rfloor}^0\right\}$ in the mask M_0 is set to 1, as a new mask $M_{d0} \in \mathbb{R}^{H_{in} \times W_{in}}$, while the value of $\left\{h_k^1, ..., h_{k+\lfloor d \times H_{in} \rfloor}^1\right\}$ is set to 0, as a new mask $M_{d1} \in \mathbb{R}^{H_{in} \times W_{in}}$. The mask feature extraction can be represented as:

$$Y_m = c^{3\times3\times3}(X_{in} \otimes M_{d0}) + c^{3\times3\times3}(X_{in} \otimes M_{d1}), \tag{2}$$

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

In this paper, we propose two ways to combine these two feature maps. One is the element-wise addition (CMCL-A), which can be designed as

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

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

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

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

As shown in Fig.3, during the test stage, we directly input the original feature map X_{in} into the mask feature extraction to generate the local feature representations.

1.2. Query Expansion

In general, to evaluate the performance, we divide the dataset into two subsets: the gallery set and the probe set. In

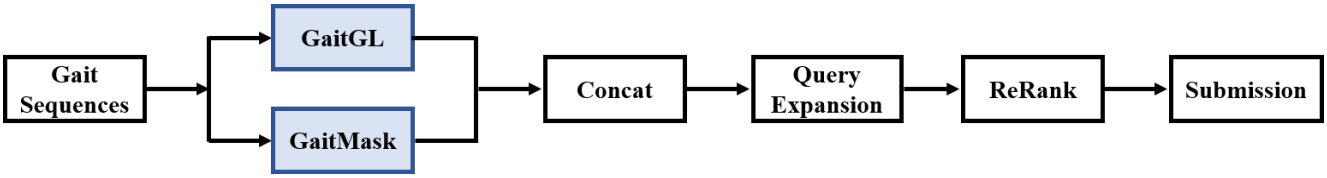

Figure 1. The inference pipeline of our framework.

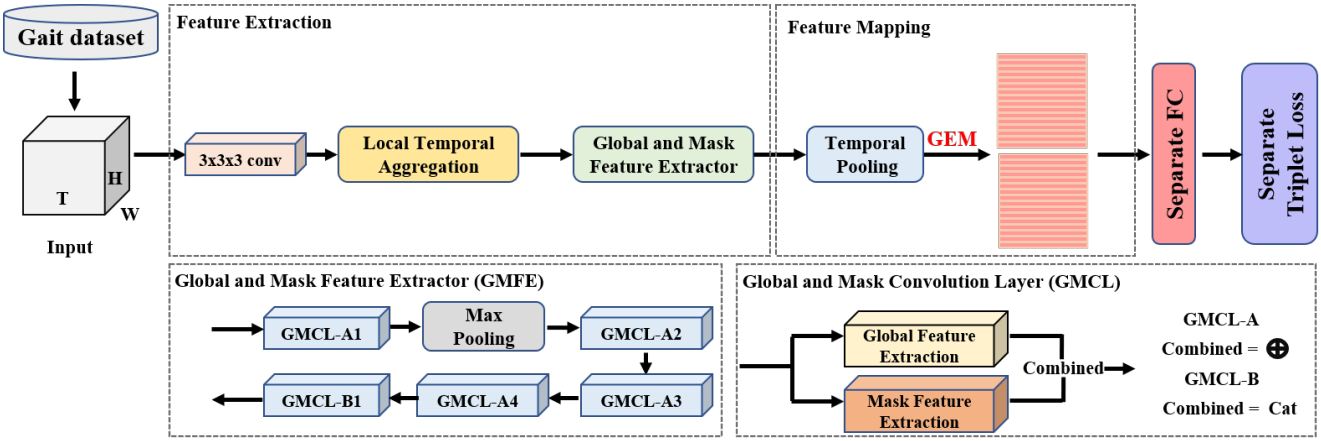

Figure 2. Overview of the proposed GaitMask method.

this paper, we adopt query expansion technology to improve the recognition accuracy. Specifically, we first concatenate all feature representations from the gallery and probe sets. Then, we adopt a clustering method based on the euclidean distance to find the most similar samples. Each feature representation from the gallery and probe sets is updated to the mean feature representation of the other representations in the same cluster.

1.3. Implementation Details

We use a preprocessing method [1] to normalize the gait data from the CASIA-E and OUMVLP datasets. The size of the normalized gait images is 64×44. The training details are shown in Tab.1. When we train two networks on CASIA-B and CASIA-E datasets, we take the trained parameters from the OUMVLP dataset as the pre-trained parameters. The network parameters of the GaitMask are shown in Table.2. All experiments take Adam as the optimizer and the initial learning rate is 1e-4. For the OUMVLP dataset, the learning rate reset to 1e-5 after 150K. For the CASIA-B and CASIA-E datasets, the learning rate reset to 1e-5 after 10K.

2. Experiments

In our work, the proposed method uses several techniques to improve the recognition accuracy, e.g. model ensemble, query expansion, and re-ranking. In this section, we design different ablation studies to evaluate the effectiveness of these techniques.

The experimental results are shown in Tab.3. It can be observed that each technique plays an important role in the whole framework.

Table 1. The training details of the proposed framework on different datasets. T means the length of input gait sequences in the training stage.

Datasets	Method	Epoch	T Frames	BatchSize
OUMVLP	GaitGL	250K	30	32*8
	GaitMask			
CASIA-B and CASIA-E	GaitGL	15K	64	12*8
	GaitMask			

Table 2. Network parameters of the proposed GaitMask

Layer Name	In_C	Out_C	Kernel	Global	N-part
First Conv	1	32	(3,3,3)	✓	×
LTA	32	32	(3,1,1)	–	–
GMCL-A1	32	64	(3,3,3)	✓	2
Max Pooling, kernel size =(1, 2, 2), stride=(1, 2, 2)					
GMCL-A2	64	128	(3,3,3)	✓	2
GMCL-A3	128	128	(3,3,3)	✓	2
GMCL-A4	128	128	(3,3,3)	✓	2
GMCL-B1	128	128	(3,3,3)	✓	2

Table 3. Rank-1 accuracy (%) of different techniques. QE means the Query Expansion technique.

Model	Re-ranking	QE	Accuracy
GaitGL			64.1%
GaitGL	✓		78.4%
GaitGL	✓	✓	80.7%
GaitGL+GaitMask	✓	✓	83.7%

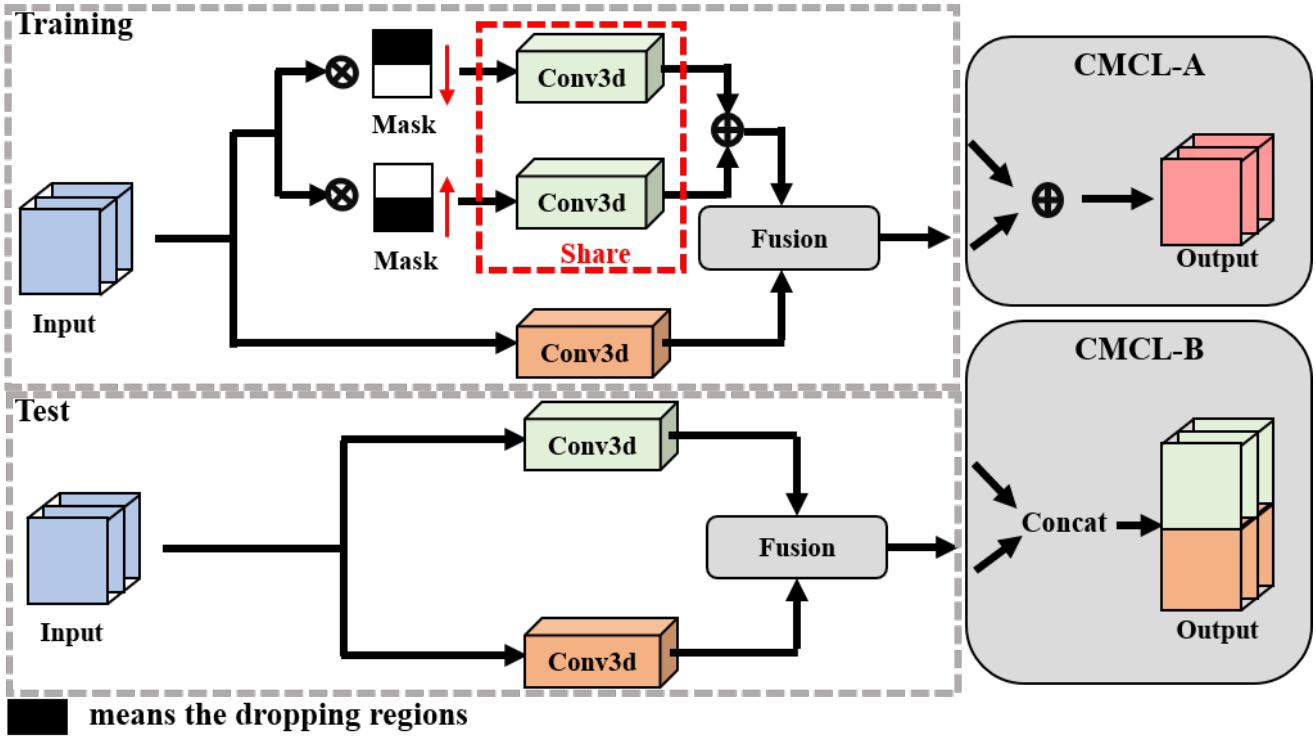

Figure 3. Overview of the proposed Global and Mask Convolution Layer.

References

[1] H. Chao, Y. He, J. Zhang, and J. Feng. Gaitset: Regarding gait as a set for cross-view gait recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 8126–8133, 2019.

[2] B. Lin, S. Zhang, and F. Bao. Gait recognition with multiple-temporal-scale 3d convolutional neural network. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3054–3062, 2020.

[3] B. Lin, S. Zhang, X. Yu, Z. Chu, and H. Zhang. Learning effective representations from global and local features for cross-view gait recognition. *arXiv preprint arXiv:2011.01461*, 2020.

[4] Z. Zhong, L. Zheng, D. Cao, and S. Li. Re-ranking person re-identification with k-reciprocal encoding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1318–1327, 2017.