

Gait Recognition with Mask-based Regularization

Chuanfu Shen^{1,2†}, Beibei Lin^{3†}, Shunli Zhang³,
George Q. Huang¹, Shiqi Yu^{2*}, and Xin Yu⁴

¹ The University of Hong Kong

² Southern University of Science and Technology

³ Beijing Jiaotong University

⁴ University of Technology Sydney

Abstract. Most gait recognition methods exploit spatial-temporal representations from static appearances and dynamic walking patterns. However, we observe that many part-based methods neglect representations at boundaries. In addition, the phenomenon of overfitting on training data is relatively common in gait recognition, which is perhaps due to insufficient data and low-informative gait silhouettes. Motivated by these observations, we propose a novel mask-based regularization method named ReverseMask. By injecting perturbation on the feature map, the proposed regularization method helps convolutional architecture learn the discriminative representations and enhances generalization. Also, we design an Inception-like ReverseMask Block, which has three branches composed of a global branch, a feature dropping branch, and a feature scaling branch. Precisely, the dropping branch can extract fine-grained representations when partial activations are zero-outed. Meanwhile, the scaling branch randomly scales the feature map, keeping structural information of activations and preventing overfitting. The plug-and-play Inception-like ReverseMask block is simple and effective to generalize networks, and it also improves the performance of many state-of-the-art methods. Extensive experiments demonstrate that the ReverseMask regularization help baseline achieves higher accuracy and better generalization. Moreover, the baseline with Inception-like Block significantly outperforms state-of-the-art methods on the two most popular datasets, CASIA-B and OUMVLP. The source code will be released.

Keywords: Gait Recognition; Regularization; Network Generalization

1 Introduction

Gait recognition [31] utilizes appearance and walking patterns as clues to identify people from images sequences. Gait recognition can achieve perception-free human identification at a distance, which is hardly achievable by other biometrics such as the face, fingerprint, iris. Nevertheless, gait recognition is still facing

* Corresponding author.

† C.S. and B.L. are co-first authors.

several challenges such as pose[24], carrying and clothing[40], ageing[19], illumination, and occlusions.

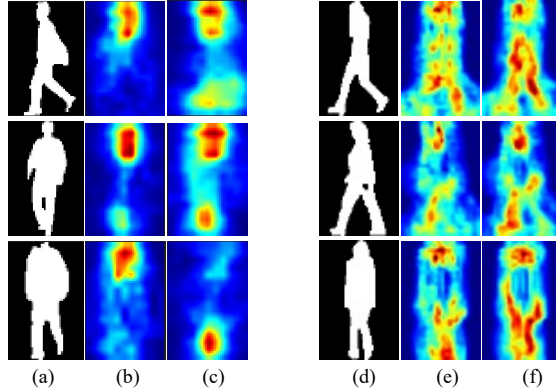


Fig. 1. Visualization [46] of activation for different methods on CASIA-B. **Left:** (a) input silhouettes. (b) activation visualization of GaitSet [4]. (c) GaitSet with our proposed ReverseMask regularization method. **Right:** (d) input silhouettes. (e) activation visualization of GaitGL [27]. (f) GaitGL with ReverseMask regularization, respectively. GaitSet concentrates on the head and foot regions. GaitGL distributes the attention on different moving parts of a human body, but the noncontinuous attention reveals that information is missing at the horizontal boundaries.

Recently, several studies [33,39,38,42,26] achieved impressive performance to facilitate gait recognition using the deep convolutional neural networks. Notably, most appearance-based gait recognition models only adopt a limited number of stacked layers (refers to depth). In contrast, other visual recognition tasks [3,13] have already greatly benefited from very deep models. However, such deep architecture is inferior to shadow networks in gait recognition. To our best knowledge, the preference to apply shallow models is mainly caused by two aspects: i) The task of gait recognition has less data for training. For example, Sports-1M and YouTube-8M [20,1] contain millions of action videos, while the largest cross-view gait dataset [35] provides 0.1 million gait sequences for training. ii) Silhouettes provide less information than information from the RGB modality. Therefore, we think that insufficient training data can easily lead many deep networks [4,15,44] to risk overfitting the salient characteristics. The *overfitting* phenomenon can be glimpsed from the activations visualization of classical GaitSet[4] as shown in Fig. 1 (b). The learned representations easily focus on the most discriminative pattern, but it leads to poor generalization performance on validation. Unexpectedly, this problem has not been indicated well yet in the previous literature.

By partitioning the holistic human body into many horizontal parts, the part-based methods [8,27,42] leverage partial features to prevent the issue of *overfitting* on salient representations. Nonetheless, the learned representations by part-based methods are noncontinuous and distributed sparsely as shown in

Fig. 1 (e). The conventional part-based methods employ hard boundary partition, where spatial clues of each part can only concentrate on inner partial regions, neglecting inter-part correlation. This phenomenon of noncontinuous representations refers to *boundary* isolation in this paper. Therefore, it is necessary to prevent *overfitting* and *boundary*, furthermore improving the generalization ability and performance of deep networks for robust gait recognition.

To address the *overfitting* issue, various data augmentation and regularization methods [34,5] have been proposed, such as input-level random erasing [45] and feature-level DropBlock [11]. The principle of these methods is to inject noise into raw input or feature, producing extra data so that convolutional networks do not overfit the training data. We argue that the main drawback of various erasing-based methods is that it only zero-out features. Besides, it also can apply perturbation by scaling activations. The scaling features extents to bring about more noise to prevent network overfitting. Moreover, it is also perfectly suitable for appearance-based gait recognition because the scaling regularization supervises the network to look for structural evidence of gait for simply silhouettes. For the *boundary* issue, we analyze this isolation is mainly caused by manual partition like GaitPart [8], where such convolutional layers can only capture internal representations but ignore semantic information at the boundary. Therefore, it is straightforward to consider generating random partition to avoid neglecting regional representations during training to overcome the *boundary* issue.

The aforementioned intuitions inspires us to address both *overfitting* and *boundary* problems by introducing a mask-based regularization method. In this work, we propose a novel regularization method called ReverseMask, with a corresponding Inception-like ReverseMask Block. Specifically, the Inception-like ReverseMask Block has a parallel architecture consisting of three branches, which are global branch, dropping branch, and scaling branch. Within both dropping branch and scaling branch, the novel ReverseMask layer is plugging as a regularizer where receiving features from the previous layer and then producing a pair of features with perturbation. Specifically, the ReverseMask layer zero-outs partial features for the dropping branch. Therefore, the dropping branch effectively captures fine-grained representations since convolutional filters must look at informative regions to fit the erased feature. For the scaling branch, ReverseMask randomly scales the value of activations so that the perturbation forces a convolutional filter to learn the structural characteristics of gait silhouettes.

In our experiments, adding Inception-like ReverseMask Block to GaitSet shows its regularizing convolutional networks as shown in Fig. 1 as well as improving generalization performance in cross-view gait recognition under clothing from 74.8% to 76.3%. Besides, Inception-like ReverseMask Block relieves the phenomena of *boundary* isolation of part-based methods and improves cross-view gait recognition accuracy as well. In summary, the contributions of this work are listed as follows:

- The novel ReverseMask layer is superior to regular DropBlock regularization in our experiments. The ReverseMask provides much more stable regularization and speeds up the training.

- We propose a novel Inception-like ReverseMask Block with the scaling branch, which helps to capture structural gait representations. In particular, Inception-like ReverseMask Block can be flexibly embedded into the most recent gait frameworks, improving the discriminativeness of feature representations and the generalization performance of models.
- The network with proposed regularization method outperforms start-of-the-art methods on two popular benchmarks: CASIA-B [40] and OUMVLP [35].

2 Related Work

2.1 Gait Recognition

Holistic-based recognition. To extract holistic gait features, template-based methods [33,39] utilize convolutional neural networks directly on gait templates like Gait Energy Images [12] (GEI) in early. Wu *et al.* [39] propose three types of architecture to recognize the most discriminative changes of gait features and provide many comprehensive experiments on cross-view gait recognition performance. Nevertheless, template-based methods lose temporal and fine-grained spatial information. In contrast, many sequence-based methods [4,15,43] conduct feature extractors on each frame to capture detailed spatial clues across frames. For example, Chao *et al.* [4] regard a gait as a set of independent frames, then aggregate and fuse the extracted set-level feature by the proposed set pooling unit. However, the gait recognition methods based on holistic representations tend to focus on the most representative salient patterns, leading to distinguishing the subjects trickily.

Part-based recognition. The part-based methods [8,27,42] extensively exploit fine-grained spatial cues from multiple parts for local representation learning. The part-based methods can extract from different types of local regions *i.e.* patch [30], body components [28,24], vertical or horizontal bins [9], and attentive regions [22]. GaitPart [8] introduces the Focal Convolution Layer to enhance the part-level spatial features by splitting the input feature map into several parts horizontally. Horizontal Pooling is widely used in many approaches [15,16,27] since GaitSet [4] adopts the Horizontal Pyramid Pooling [10] from person re-identification. Specifically, Horizontal Pyramid Pooling separates feature maps into hierarchical strips with multi-scales hierarchy, improving spatial discriminative ability. Most works [27,8,42] are based on pre-defined uniform partition leading to the problem of boundary. Zhang *et al.* [42] propose a method based on the learned partition, but boundary isolation still exists. Recently, GaitGL [27] combines both global and partial features, gaining more robust spatial representation. Besides, it also introduces 3D ConvNets to learn spatial-temporal integrated features for gait recognition. However, no matter learned or hand-crafted partition-based methods have introduced the boundary problem, neglecting the spatial information at the foundation.

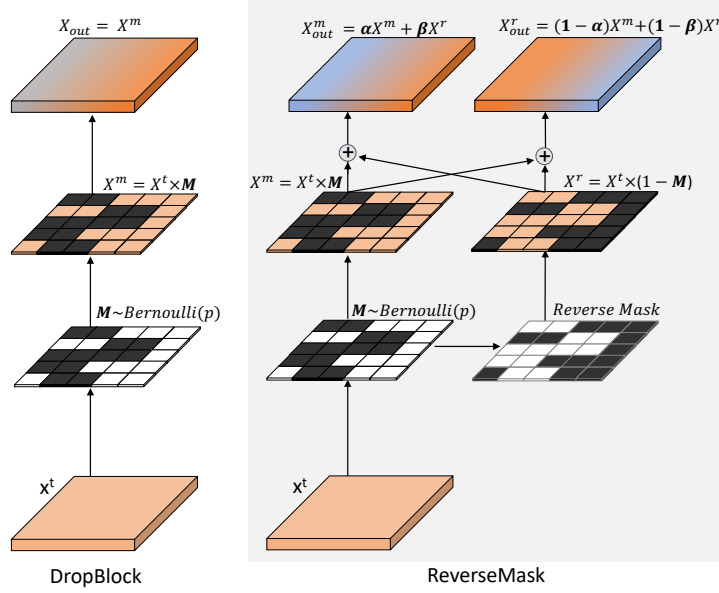


Fig. 2. Illustration of our proposed ReverseMask. Given two random variable, ReverseMask introduces perturbation on networks by reducing values of feature map on selected regions.

2.2 Erasing Images or Activations

Image based erasing [6,45] is widely adopted as a data augmentation technique. In recent years, cutout [6] has been demonstrated that masking out partial feature maps can improve generalization of convolutional neural networks and achieve better performance in many tasks such as object detection [2] and person re-identification [45].

Feature based erasing [5,11] is an alternative regularization technique that is implemented by using zero-masking directly on the feature map. Dropout [34] is effective to prevent overfitting, but it designs initially for fully-connected layer. While the mechanism of dropout also brings many successful works on convolutional neural networks, such as SpatialDropout [37] and DropBlock [11].

Our method drops identical, randomly selected regions of convolutional features for sequences in a batch, which has been proved effective in previous literature [5]. However, our work presents a scaling mechanism where parts of convolutional features are multiplied by a random ratio [36]. Feature maps are multiplied by a random ratio rather than zero, presenting novel structural representations for robust gait recognition. Furthermore, we also argue that such structural representations are ignored, while many methods [8,25] only proposed to capture local representations over coarse pre-defined parts[27,9].

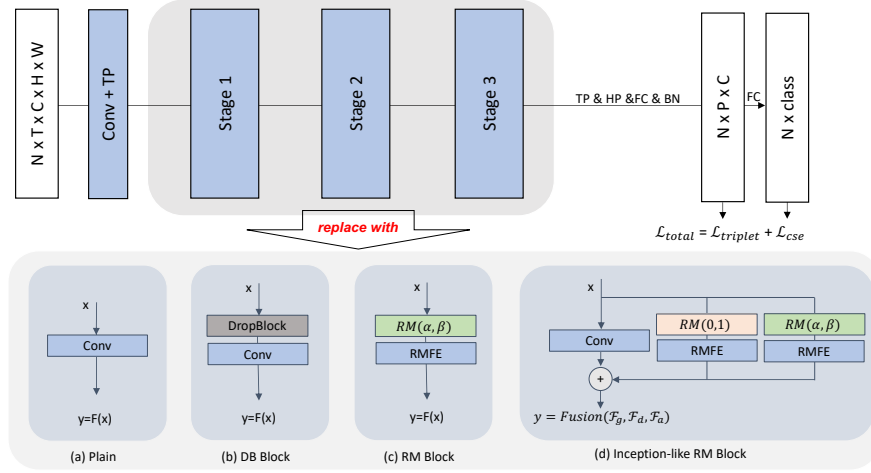


Fig. 3. The framework of our proposed method. **Top:** a simplified model [27] as our baseline model; **Bottom:** Four replaceable blocks. *TP*, *FC*, *HP* and *BN* mean temporal pooling aggregation layer [27,4], fully connected layer, horizontal pooling layer[10], and batch normalization layer [18], respectively. *Conv* indicates a convolutional layer followed by *Leaky ReLU* [29]. ***RM*** is our proposed *ReverseMask* layer, and it receive a pair variable to inject noise on feature map. While ***RMFE*** is the feature extraction layer specifically designed for *ReverseMask*.

3 Our approach

In this work, we propose a novel regularization method named ReverseMask, which can help relieve convolutional architecture overfitting on training sets. In this section, we first formulate the simple yet effective ReverseMask (Sec. 3.1). Then the detailed definitions of other corresponding components are followed (Sec. 3.2). Finally, two variants of building block are introduced (Sec. 3.3) for making ReverseMask as easy-to-use as possible.

3.1 ReverseMask Layer

ReverseMask is a simple regularization method similar to DropBlock. Its main difference from DropBlock is significant. (1) ReverseMask is a general mask-based regularizer, which produces a pair of masked features instead of only one masked feature. (2) ReverseMask introduces perturbation on networks by reducing values of feature maps on selected regions. Illustrations of ReverseMask and DropBlock are shown in Fig. 2. Notably, there are only two parameters for ReverseMask, which are *reg_prob* and *p*. *reg_prob* determines the probability of changing activations undergoing ReverseMask. *p* controls the area ratio of Masked generation.

During the training stage, we first randomly sample mask on the given feature map $X_l \in \mathbb{R}^{b_l \times n_l \times c_l \times h_l \times w_l}$ at the *l*th layer. Then its paired mask can be

determined by reversing the mask matrix. The process can be defined as

$$M_{i,j}^m \sim \text{Bernoulli}(p) \quad (1)$$

$$M_{i,j}^p = 1 - M_{i,j}^m \quad (2)$$

where $M^m \in \mathbb{R}^{h_l \times w_l}$ and $M^p \in \mathbb{R}^{h_l \times w_l}$ refer to sampled mask and its paired mask. Notably, such a pair of masks is temporally synchronized, and we only conduct experiments with shared masks across different feature channels for simplicity and fair comparison with part-based methods. Then, applying the paired of reversal masks on given features, two masked features can be obtained

$$X_l^m = X_l \times M^m \quad (3)$$

$$X_l^r = X_l \times M^r \quad (4)$$

It is worth noticing that these masked features are identical to features produced by DropBlock, but ReverseMask has a pair of masked features. Finally, we introduce perturbation on activations for regularization with a pair of randomly generated values

$$\alpha, \beta \sim \text{Uniform}(0, 1) \quad (5)$$

$$X_{l_{out}}^m = \alpha X_l^m + \beta X_l^r \quad (6)$$

$$X_{l_{out}}^p = (1 - \alpha)X_l^m + (1 - \beta)X_l^r \quad (7)$$

where two random variables α and β satisfy the uniform distribution to perturb activations. $X_{l_{out}}^m \in \mathbb{R}^{b_l \times n_l \times c_l \times h_l \times w_l}$ and $X_{l_{out}}^p \in \mathbb{R}^{b_l \times n_l \times c_l \times h_l \times w_l}$ are obtained as input of further feature learning. Therefore, our proposed ReverseMask is a more general DropBlock, generating zero-masked features and scaling features to regularize the network. Beyond feature-level erasing methods, our proposed mask-based scaling regularization method is able to capture structural information from gait silhouettes. At the same time, erasing-based regularizer tends to perform better on fine-grained representation learning.

Mask Sampling. In the experiments, we have studied many mask sampling strategies, such as masking independent random units as shown in Fig. 2, and masking continuous regions. The detailed analysis refers to Sec. 4.3.

Setting the value of *reg_prob*. In our implements, *reg_prob* = 1 is constant for fair comparison with part-based methods. Taking GaitGL as an example, it can be seen as mask-based with special dropping out certain regions of activations.

Setting the value of p . We investigate the area ratio for the masked region in the experiments. p is applied to 0.5 as the optimal ratio, which means half of the feature map tends to be perturbed.

Setting the value of α and β . The ReverseMask is a flexible and general regularization method. By setting the different values of α and β , the networks regularized by ReverseMask could perform extremely differently in representation learning. We found that ReverseMask enables a network to capture fine-grained representations, like conventional part-based methods doing. The ReverseMask resembles DropBlock when $\alpha = 1, \beta = 0$. When α, β are sampled from the uniform distribution, information about the structure can still be sent to further extraction. In summary, there are two variants of ReverseMask according to the value of α, β . We illustrate these two variants of ReverseMask in different colors as shown in figure 2.

3.2 ReverseMask Feature Extraction

The conventional part-based methods utilize a shared convolutional layer to extract local representations. While GaitPart [8] utilized two-dimensional filters, and GaitGL [27] adopted three-dimensional convolutions which enhanced spatiotemporal feature learning. In other words, focal convolutions apply to each part with shared parameters. The receptive field of the focal convolution layer is restricted, which leads to the issue of *boundary*. To alleviate such boundary weakness, our proposed ReverseMask randomly generates a pair of zero-masking features when $\alpha = 1, \beta = 0$. Therefore, the convolutions can still capture the fine-grained representations without boundary neglect. In our implements, we use identical 3D convolution to GaitGL[27] for a fair comparison with part-based methods, which can design as

$$F_l = W(X_{l_{out}}^m) + W(X_{l_{out}}^p) \quad (8)$$

where $W(\cdot)$ means a 3D convolution operation, and $X_{l_{out}}^m$ and $X_{l_{out}}^p$ are the paired feature generated by Eqn. 6 and Eqn. 7. As we see, the ReverseMask Feature Extraction layer remains a full feature map for further layers, which brings about many advantages.

3.3 Variants of Building Block

Plain ReverseMask Block is a plain-like block similar to plain architecture as shown in Fig. 2. The Plain ReverseMask Block is general. It can resemble DropBlock-like regularization when $\alpha = 1, \beta = 0$, which refers to dropping Plain-RMB for short. Also, it resembles scaling regularization when α, β sample from the uniform distribution, which refers to scaling Plain-RMB for short. In our experiments, both scaling and dropping Plain-RMB are superior to DropBlock.

Inception-like ReverseMask Block is designed for two reasons: (1) It is necessary to design such multi-branches architecture to establish a fair comparison with GaitGL. (2) The ReverseMask with different settings changes the distribution of the training set. Therefore such Inception-like ReverseMask Block can take advantage of multiple branches. As shown in Fig.3(d), three representations learned from each branch are then aggregated by feature fusion module, which denotes as

$$\mathbf{F} = F_g + F_d + F_s \quad \mathbf{F} \in \mathbb{R}^{b_l \times n_l \times c_l^{out} \times h_l \times w_l} \quad (9)$$

where F_g, F_d , and F_s represent features obtained from global, dropping and scaling branch, respectively. In addition, this strategy is used in every stage, excluding the last one. The final representations are concatenated in the final stage, which is commonly used in [27,16,15]. The fusion module is represented as:

$$\mathbf{F} = \text{concat}(F_g, F_d, F_s) \quad \mathbf{F} \in \mathbb{R}^{b_l \times n_l \times c_l^{out} \times 3h_l \times w_l} \quad (10)$$

4 Experiments

The proposed method is evaluated on two popular public datasets: CASIA-B and OU-MVLP. CASIA-B is easy to evaluate the robustness to different variations, and OU-MVLP is the largest public gait dataset. Implementation details, results, comparisons, ablation study, and analysis are presented in the following part of this section.

4.1 Settings

Datasets. *CASIA-B* [40] consists of 124 subjects with ten sequences under 11 views, resulting in $124 \times 10 \times 11 = 13640$ gait sequences. Each subject walked ten times in three conditions, *i.e.* six in normal (NM), two with a bag (BG), two with a coat (CL). To compare fairly, we follow the experiment protocol in [4] which is widely employed by many other methods. Our experiments are conducted on three different configurations: Small-scale Training (ST), Medium-scale Training (MT), and Large-scale Training (LT). CASIA-B is split into a training set with 24 subjects and a test set with 100 subjects in ST configuration. MT configuration split CASIA-B to a training set with 62 subjects and a test set with 62 subjects. LT configuration has 74 subjects for training and 50 subjects for testing. Each subject's first four sequences (NM#01-NM#04) are put into the gallery set in the test phase. The remaining two sequences of NM, BG, and CL are in three different probe sets respectively to evaluate the robustness to different variations.

OU-MVLP [35] was created by Osaka University and is the largest public gait database. It contains 10307 subjects, and each subject walks twice under 14 views. So, there are $2 \times 14 = 28$ sequences for each subject. In our experiments, we follow the same protocol used in [4] also. That means 5153 subjects in the training set and the other 5154 subjects in the test set. In the test phase, sequence NM#-01 is put into the gallery set, and sequence NM#-00 is in the probe set.

Implementation details: **(1)** Human body silhouettes are aligned, cropped and resized to 64×44 by the preprocessing method [35]. The sequence length is 30 frames in the training phase, while the whole sequence is used in the test phase. **(2)** The separate Batch All (*BA+*) triplet loss [14] is applied to train our model. The batch size (p, k) is set up as (8, 16) for CASIA-B and (32, 8) for OU-MVLP, respectively. Margin m for triplet loss is set to 0.2. **(3)** The pooling parameter p is set to 6.5 for Generalized-Mean pooling [32]. **(4)** The iteration is set to 60K, 80K, and 80K for ST, MT, and LT, respectively, in CASIA-B training. In OU-MVLP training, the iteration is increased to 210K since there is more data in OU-MVLP. **(5)** Adam [21] is employed as the optimizer, and the initial learning rate is $1e-4$ with weight decay $5e-4$. The learning rate reduces to $1e-5$ after 70K iterations for MT and LT, while the learning rate changes to $1e-5$ after 150K, then $5e-6$ after 200K iterations for OU-MVLP. **(6)** For OU-MVLP, we doubled channel sizes of all convolution layers since OU-MVLP has more data than in CASIA-B.

Evaluation metric Rank-1 accuracy excluding identical-view sequences is taken as the same evaluation metric used in [4] and some other state-of-the-art methods. The metric has been widely employed to evaluate the performance of cross-view gait recognition.

4.2 Comparisons with State-of-the-Art

Results on CASIA-B. Tab. 1 shows our performance on CASIA-B, with three training configurations. With the large-scale configuration, we obtain 97.7%, 95.3%, and 86.0% on rank-1 accuracy under NM, BG, and CL, respectively. The performance of our proposed method surpasses the classical method GaitSet [4] by a large margin. Compared with two typical part-based methods, GaitPart [8] and GaitGL [27], our model outperforms GaitPart by 7.3% under the most challenging condition of clothing (CL). At the same time, it outperforms the significant accuracy of GaitGL with 2.4% under CL. The comparison with part-based methods replied on the coarse partition demonstrated Inception-like ReverseMask Block’s effectiveness in enhancing the model by integrating structural and fine-grained representations. A recent part-based method, 3DLocal [17], models the dynamic motion information on learned partition differently from previous pre-defined parts. However, the result illustrates that our model outperforms 3DLocal under all walking conditions significantly, which shows that random partition with simpleness can compete with adaptive region localization. We analyze the boosting performance coming from two aspects: (1) The proposed ReverseMask regularization injects noise into the feature map, which helps to prevent models from overfitting training data. Therefore, a better generalization performance is obtained in the test dataset. (2) random mask sampling considers full local representations from silhouettes, while the conventional part-based methods neglect characteristics around the gap between horizontal stripes. Therefore, the mask-based model should be superior to the previous part-based method depending on fixed horizontal stripes.

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

| Gallery NM#1-4 | | | 0°-180° | | | | | | | | | | | |
|----------------|--------|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Probe | | | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | Mean |
| ST (24) | NM#5-6 | GaitSet [4] | 64.6 | 83.3 | 90.4 | 86.5 | 80.2 | 75.5 | 80.3 | 86.0 | 87.1 | 81.4 | 59.6 | 79.5 |
| | | GaitGL [27] | 77.0 | 87.8 | 93.9 | 92.7 | 83.9 | 78.7 | 84.7 | 91.5 | 92.5 | 89.3 | 74.4 | 86.0 |
| | | Ours | 78.1 | 89.2 | 95.3 | 92.6 | 83.9 | 79.7 | 85.2 | 91.8 | 93.2 | 89.6 | 73.9 | 86.6 |
| | BG#1-2 | GaitSet [4] | 55.8 | 70.5 | 76.9 | 75.5 | 69.7 | 63.4 | 68.0 | 75.8 | 76.2 | 70.7 | 52.5 | 68.6 |
| | | GaitGL [27] | 68.1 | 81.2 | 87.7 | 84.9 | 76.3 | 70.5 | 76.1 | 84.5 | 87.0 | 83.6 | 65.0 | 78.6 |
| | | Ours | 70.6 | 81.7 | 88.9 | 86.9 | 76.3 | 71.1 | 77.6 | 85.7 | 88.8 | 83.8 | 67.2 | 79.9 |
| | CL#1-2 | GaitSet [4] | 29.4 | 43.1 | 49.5 | 48.7 | 42.3 | 40.3 | 44.9 | 47.4 | 43.0 | 35.7 | 25.6 | 40.9 |
| | | GaitGL [27] | 46.9 | 58.7 | 66.6 | 65.4 | 58.3 | 54.1 | 59.5 | 62.7 | 61.3 | 57.1 | 40.6 | 57.4 |
| | | Ours | 50.2 | 65.4 | 70.8 | 69.0 | 63.0 | 58.0 | 63.3 | 67.6 | 66.2 | 61.6 | 43.2 | 61.7 |
| | NM#5-6 | GaitSet [4] | 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 [27] | 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 | 93.6 | 98.1 | 98.8 | 97.7 | 95.6 | 93.6 | 95.9 | 98.8 | 98.8 | 96.9 | 92.1 | 96.4 |
| MT (62) | BG#1-2 | GaitSet [4] | 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 [27] | 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 | 89.6 | 95.5 | 96.8 | 95.5 | 92.2 | 87.0 | 90.9 | 95.5 | 98.2 | 94.6 | 87.1 | 93.0 |
| | CL#1-2 | GaitSet [4] | 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 [27] | 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 | 73.1 | 85.9 | 90.6 | 88.4 | 80.6 | 75.5 | 81.5 | 86.5 | 87.4 | 81.4 | 66.5 | 81.6 |
| | NM#5-6 | GaitSet [4] | 90.8 | 97.9 | 99.4 | 96.9 | 93.6 | 91.7 | 95.0 | 97.8 | 98.9 | 96.8 | 85.8 | 95.0 |
| | | GaitPart [8] | 94.1 | 98.6 | 99.3 | 98.5 | 94.0 | 92.3 | 95.9 | 98.4 | 99.2 | 97.8 | 90.4 | 96.2 |
| | | GaitGL [27] | 96.0 | 98.3 | 99.0 | 97.9 | 96.9 | 95.4 | 97.0 | 98.9 | 99.3 | 98.8 | 94.0 | 97.4 |
| | | 3DLocal [17] | 96.0 | 99.0 | 99.5 | 98.9 | 97.1 | 94.2 | 96.3 | 99.0 | 98.8 | 98.5 | 95.2 | 97.5 |
| | | Ours | 96.5 | 98.4 | 99.2 | 98.0 | 97.1 | 95.5 | 97.4 | 99.2 | 99.3 | 99.1 | 95.0 | 97.7 |
| | | GaitSet [4] | 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
| LT (74) | BG#1-2 | GaitPart [8] | 89.1 | 94.8 | 96.7 | 95.1 | 88.3 | 84.9 | 89.0 | 93.5 | 96.1 | 93.8 | 85.8 | 91.5 |
| | | GaitGL [27] | 92.6 | 96.6 | 96.8 | 95.5 | 93.5 | 89.3 | 92.2 | 96.5 | 98.2 | 96.9 | 91.5 | 94.5 |
| | | 3DLocal [17] | 92.9 | 95.9 | 97.8 | 96.2 | 93.0 | 87.8 | 92.7 | 96.3 | 97.9 | 98.0 | 88.5 | 94.3 |
| | | Ours | 93.7 | 97.0 | 97.3 | 95.8 | 94.9 | 91.4 | 93.5 | 97.3 | 98.3 | 97.3 | 92.4 | 95.3 |
| | CL#1-2 | GaitSet [4] | 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 [8] | 70.7 | 85.5 | 86.9 | 83.3 | 77.1 | 72.5 | 76.9 | 82.2 | 83.8 | 80.2 | 66.5 | 78.7 |
| | | GaitGL [27] | 76.6 | 90.0 | 90.3 | 87.1 | 84.5 | 79.0 | 84.1 | 87.0 | 87.3 | 84.4 | 69.5 | 83.6 |
| | | 3DLocal [17] | 78.2 | 90.2 | 92.0 | 87.1 | 83.0 | 76.8 | 83.1 | 86.6 | 86.8 | 84.1 | 70.9 | 83.7 |
| | | Ours | 78.9 | 91.5 | 93.1 | 91.1 | 85.6 | 81.0 | 85.2 | 89.0 | 90.9 | 87.3 | 72.9 | 86.0 |
| | | GaitSet [4] | 61.4 | 75.4 | 80.7 | 77.3 | 72.1 | 70.1 | 71.5 | 73.5 | 73.5 | 68.4 | 50.0 | 70.4 |

Results on OU-MVLP. Tab. 2 lists the rank-1 accuracy of our model and other state-of-the-art methods on OU-MVLP. Our method achieves the best performance of 90.9% on cross-view gait recognition. It noticed that 3DLocal obtains equal accuracy to our method when the invalid probe sequences are included, but it only demonstrates its performance of 96.5% if the invalid probe sequences are excluded. Our method achieves 97.5% rank-1 accuracy, a much better result.

4.3 Ablation study

Masking strategies. We study five variants of mask sampling strategies. The illustration and implements are detailed in supplementary materials. Those variants of sampling strategies achieve comparable performance, while different sampling strategies impact a lot on performance in other visual recognition tasks. We analyze that other visual tasks can take advantage of RGB modality, while silhouette-based gait recognition tends to capture static features and motion patterns from low-informative data.

Table 2. Rank-1 accuracy (%) on OUMVLP dataset under different view angles, excluding identical-view cases. The top eight rows and bottom six rows show the results with and without invalid probe sequences, respectively.

| Method | Probe View | | | | | | | | | | | | | | | Mean |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|------|
| | 0° | 15° | 30° | 45° | 60° | 75° | 90° | 180° | 195° | 210° | 225° | 240° | 255° | 270° | | |
| GaitSet [4] | 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 [8] | 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 [15] | 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 [41] | 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 | |
| GaitGL [27] | 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 | |
| CSTL [16] | 87.1 | 91.0 | 91.5 | 91.8 | 90.6 | 90.8 | 90.6 | 89.4 | 90.2 | 90.5 | 90.7 | 89.8 | 90.0 | 89.4 | 90.2 | |
| 3DLocal [17] | 86.1 | 91.2 | 92.6 | 92.9 | 92.2 | 91.3 | 91.1 | 86.9 | 90.8 | 92.2 | 92.3 | 91.3 | 91.1 | 90.2 | 90.9 | |
| Ours | 87.9 | 91.5 | 91.7 | 92.0 | 92.0 | 91.6 | 91.3 | 90.7 | 90.3 | 90.9 | 91.1 | 90.8 | 90.5 | 90.2 | 90.9 | |

| | | | | | | | | | | | | | | | |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| GaitSet [4] | 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 [8] | 88.0 | 94.7 | 97.7 | 97.6 | 95.5 | 96.6 | 96.2 | 90.6 | 94.2 | 97.2 | 97.1 | 95.1 | 96.0 | 95.0 | 95.1 |
| GLN [15] | 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 |
| GaitGL [27] | 90.5 | 96.1 | 98.0 | 98.1 | 97.0 | 97.6 | 97.1 | 94.2 | 94.9 | 97.4 | 97.4 | 95.7 | 96.5 | 95.7 | 96.2 |
| 3DLocal [17] | - | - | - | - | - | - | - | - | - | - | - | - | - | - | 96.5 |
| Ours | 93.7 | 97.5 | 98.6 | 98.8 | 98.0 | 98.5 | 98.2 | 96.5 | 96.7 | 98.2 | 98.1 | 97.1 | 97.6 | 97.2 | 97.5 |

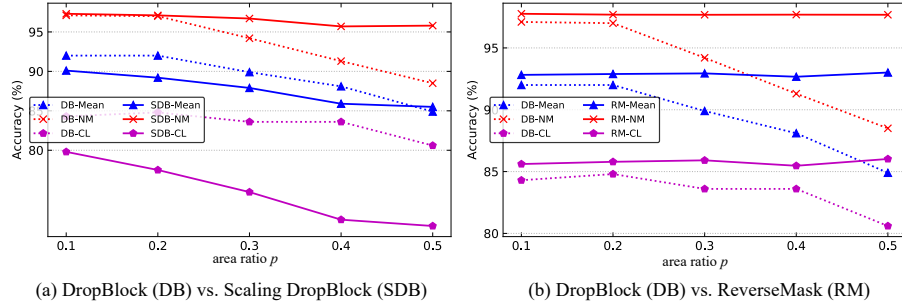


Fig. 4. Performance comparison with different area ratio for sampling selected masking.

Impact of masked area ratio p . As Fig. 4 shows, we found the stacked structure of DropBlock with the harsh setting of feature dropping can result in performance degradation. To our best knowledge, these results can be explained in two ways. (1) The harsh regularization changes data distribution dramatically; (2) The feature dropping in the shallow layer provides the incomplete feature map for the following layer, obstructing the following convolutions to learn representations.

Benefit of ReverseMask. According to the analysis of performance degradation when harshly dropping feature map, we design ReverseMask Feature Extraction, which compensates the incomplete feature by its reversal masked feature. Albeit, the ReverseMask is simple, our experiments in Fig. 4(b) show that it is effective to alleviate the issue of performance degradation led by conventional DropBlock.

Mask-based vs. Part-based. Our proposed mask-based regularizer can resemble part-based when the mask is constant instead of randomly generated. To illustrate the transition from mask-based model to conventional part-based model, we demonstrate such transition in the supplementary materials. The results in Tab. 3 indicate that the mask-based model can outperform the part-based model by utilizing the information at the boundary, which is neglected by part-based methods.

Table 3. Gait recognition performance reported in *rank-1* accuracy on CASIA-B.

| Model | Dim | NM | BG | CL | Mean |
|---|-----------------|-------------|-------------|-------------|-------------|
| Baseline [27] | 4096 | 97.0 | 93.9 | 83.3 | 91.4 |
| Baseline + part-based [27] | 4096×2 | 97.4 | 94.5 | 83.8 | 92.0 |
| Baseline + Inception RMB(<i>w/o</i> dropping branch) | 4096×2 | 97.6 | 94.8 | 85.1 | 92.5 |
| Baseline + Inception RMB(<i>w/o</i> scaling branch) | 4096×2 | 97.5 | 94.7 | 85.2 | 92.5 |
| Baseline + Inception RMB(<i>w/o</i> global branch) | 4096×2 | 97.6 | 95.0 | 85.1 | 92.6 |
| Baseline + Inception RMB | 4096×3 | 97.7 | 95.3 | 86.0 | 93.0 |

Feature scaling vs. Feature dropping. As shown in Fig.4(a), the curve reveals two significant phenomenon: (1) Although the trends of performance are the same for both scaling and dropping regularization, the model performs robust in the condition of clothing setting when dropping regularization is applied. (2) The model with scaling regularization obtains better accuracy on both normal and bag-carrying conditions. In the literature[23,43], gait structural representations contribute to distinguish subjects especially for normal and bag-carrying conditions, since the appearance information are relatively complete in these two conditions. It conducts that feature scaling regularizes model to learn structural representations which is robust on the conditions of normal wearing and bag-carrying, and feature dropping regularization enforces the networks to look at fine-grained features which is robust on the clothing condition.

Variants of Building Block. Two variants of blocks are built for evaluating proposed ReverseMask regularization method. As mentioned previously, Plain ReverseMask Block although has not improved performance to baseline model, but it helps model alleviate the drawback of DropBlock and prevents network from performance degradation. Besides, Scaling Plain ReverseMask Block achieves better performance than Dropping Plain ReverseMask Block. Notably, the most significant improvement is obtained when Inception-like ReverseMask Block is used to aggregate global, structural, and local representations. In addition, we think that the Plain ReverseMask Block is still with potential superiority to baseline model since the *reg_prob* set to one. In other words, it means this regularization configuration only supervises model to fix scaling data without any original silhouettes or features.

Extend to other methods. The proposed Inception-like ReverseMask Block is not only effective, but also plug-and-play. The ReverseMask regularization can generalize to the majority of gait methods, We study the extensive ReverseMask

to three representative models. In Tab.4, all models gain performance improvement after integrating ReverseMask regularizer to enhance discriminativeness of representations. We can observe that the model based on the Conv3D can gain relatively higher improvements. We analysis it is because Conv3D enhances models ability by superiority spatio-temporal representations learning.

Table 4. Effectiveness of ReverseMask regularization. To be noticed that the results of GaitSet are reproduced by [7].

| Model | NM | BG | CL | Mean |
|---|---------------------|---------------------|---------------------|---------------------|
| OpenGait | 96.3 | 92.2 | 77.6 | 88.7 |
| OpenGait + Inception-like ReverseMask Block | 97.0 (0.7) | 92.2 (0.0) | 79.4 (1.8) | 89.5 (0.8) |
| GaitSet | 95.9 | 91.3 | 74.8 | 87.3 |
| GaitSet + Inception-like ReverseMask Block | 96.1 (0.2) | 91.3 (0.0) | 76.3 (1.5) | 87.9 (0.6) |
| Baseline | 97.0 | 93.9 | 83.3 | 91.4 |
| Baseline + Inception-like ReverseMask Block | 97.7 (0.7) | 95.3 (1.4) | 86.0 (2.7) | 93.0 (1.6) |

Comparison to other regularization techniques. We compare ReverseMask to random erasing and DropBlock, which are two commonly used regularization techniques. In Tab. 5, Inception-like ReverseMask Block has better performance than not only feature-level but also input-level erasing regularization methods. Besides, we train baseline model with Inception-like ReverseMask Block and random erasing, and it achieves the best performance which shows our proposed feature-level regularization method can complementary to other regularization techniques.

Table 5. Performance of different regularization techniques. **RMB** donates ReverseMask Block for short. While Scaling DropBlock is a DropBlock-like regularization, Scaling DropBlock scales feature map rather than dropping.

| Model | NM | BG | CL | Mean |
|---|------|------|------|------|
| Baseline | 97.0 | 93.9 | 83.3 | 91.4 |
| Baseline + Random Erasing($reg_prob=0.5$) | 97.8 | 94.9 | 83.9 | 92.2 |
| Baseline + DropBlock($reg_prob=0.5, p=0.5$) | 96.2 | 93.3 | 84.5 | 91.3 |
| Baseline + DropBlock($reg_prob=1, p=0.5$) | 88.5 | 85.6 | 80.6 | 84.9 |
| Baseline + Scaling DropBlock($reg_prob=1, p=0.5$) | 95.8 | 90.3 | 70.4 | 85.5 |
| Baseline + Dropping Plain RMB($reg_prob=1$) | 96.7 | 93.3 | 82.6 | 90.9 |
| Baseline + Scaling Plain RMB($reg_prob=1$) | 97.1 | 94.1 | 83.0 | 91.4 |
| Baseline + Inception RMB($reg_prob=1$) | 97.7 | 95.3 | 86.0 | 93.0 |
| Baseline + Inception RMB($reg_prob=1$) + Random Erasing($reg_prob=0.5$) | 98.1 | 96.0 | 86.9 | 93.7 |

5 Conclusion

In this paper, we propose a novel mask-based regularization method called ReverseMask for obtaining better generalization performance, while it also vanishes the problems of overfitting and boundary existing in previous methods. To generalize the ReverseMask to the majority of works, we present the Inception-like ReverseMask Block which is able to capture more discriminative representations. In particular, feature scaling branch tends to extract structural information from silhouettes appearance, and feature dropping effectively utilizes local representations. Extensive experiments verify that the proposed regularization method achieves appealing performance on the CASIA-B and OU-MVLP. The ReverseMask feature is potentially extended to other tasks as an effective regularizer for better generalization of model.

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark. arXiv (2016)
2. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv (2020)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
4. Chao, H., He, Y., Zhang, J., Feng, J.: Gaitset: Regarding gait as a set for cross-view gait recognition. In: AAAI (2019)
5. Dai, Z., Chen, M., Gu, X., Zhu, S., Tan, P.: Batch dropblock network for person re-identification and beyond. In: ICCV (2019)
6. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv (2017)
7. Fan, C., Liang, J., Shen, C., Yu, S.: Opengait: A strong baseline and bag of tricks (2021)
8. Fan, C., Peng, Y., Cao, C., Liu, X., Hou, S., Chi, J., Huang, Y., Li, Q., He, Z.: Gaitpart: Temporal part-based model for gait recognition. In: CVPR (2020)
9. Foster, J.P., Nixon, M.S., Prugel-Bennett, A.: New area based metrics for gait recognition. In: International Conference on Audio-and Video-Based Biometric Person Authentication (2001)
10. Fu, Y., Wei, Y., Zhou, Y., Shi, H., Huang, G., Wang, X., Yao, Z., Huang, T.: Horizontal pyramid matching for person re-identification. In: AAAI (2019)
11. Ghiasi, G., Lin, T.Y., Le, Q.V.: Dropblock: A regularization method for convolutional networks. NeurIPS (2018)
12. Han, J., Bhanu, B.: Individual recognition using gait energy image. TPAMI (2005)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners (2021)
14. Hermans, A., Beyer, L., Leibe, B.: In defense of the triplet loss for person re-identification. arXiv preprint arXiv:1703.07737 (2017)
15. Hou, S., Cao, C., Liu, X., Huang, Y.: Gait lateral network: Learning discriminative and compact representations for gait recognition. In: ECCV (2020)
16. Huang, X., Zhu, D., Wang, H., Wang, X., Yang, B., He, B., Liu, W., Feng, B.: Context-sensitive temporal feature learning for gait recognition. In: ICCV (2021)
17. Huang, Z., Xue, D., Shen, X., Tian, X., Li, H., Huang, J., Hua, X.S.: 3d local convolutional neural networks for gait recognition. In: ICCV (2021)
18. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
19. Iwama, H., Okumura, M., Makihara, Y., Yagi, Y.: The ou-isir gait database comprising the large population dataset and performance evaluation of gait recognition. TIFS (2012)
20. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
21. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
22. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2018)

23. Li, X., Makihara, Y., Xu, C., Yagi, Y., Yu, S., Ren, M.: End-to-end model-based gait recognition. In: ACCV (2020)
24. Liao, R., Cao, C., Garcia, E.B., Yu, S., Huang, Y.: Pose-based temporal-spatial network (ptsn) for gait recognition with carrying and clothing variations. In: CCBR (2017)
25. Lin, B., Liu, Y., Zhang, S.: Gaitmask: Mask-based model for gait recognition. In: BMVC (2021)
26. Lin, B., Zhang, S., Bao, F.: Gait recognition with multiple-temporal-scale 3d convolutional neural network. In: ACM MM (2020)
27. Lin, B., Zhang, S., Yu, X.: Gait recognition via effective global-local feature representation and local temporal aggregation. In: ICCV (2021)
28. Liu, Z., Malave, L., Osuntogun, A., Sudhakar, P., Sarkar, S.: Toward understanding the limits of gait recognition. In: Biometric Technology for Human Identification (2004)
29. Maas, A.L., Hannun, A.Y., Ng, A.Y., et al.: Rectifier nonlinearities improve neural network acoustic models. In: ICML (2013)
30. Nandy, A., Chakraborty, R., Chakraborty, P.: Cloth invariant gait recognition using pooled segmented statistical features. Neurocomputing (2016)
31. Nixon, M.S., Tan, T., Chellappa, R.: Human identification based on gait. Springer Science & Business Media (2010)
32. Radenović, F., Tolias, G., Chum, O.: Fine-tuning cnn image retrieval with no human annotation. TPAMI (2018)
33. Shiraga, K., Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: Geinet: View-invariant gait recognition using a convolutional neural network. In: ICB (2016)
34. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR (2014)
35. 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. TCVA (2018)
36. Tang, Y., Wang, Y., Xu, Y., Shi, B., Xu, C., Xu, C., Xu, C.: Beyond dropout: Feature map distortion to regularize deep neural networks. In: AAAI (2020)
37. Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C.: Efficient object localization using convolutional networks. In: CVPR (2015)
38. Wolf, T., Babaei, M., Rigoll, G.: Multi-view gait recognition using 3d convolutional neural networks. In: ICIP (2016)
39. Wu, Z., Huang, Y., Wang, L., Wang, X., Tan, T.: A comprehensive study on cross-view gait based human identification with deep cnns. TPAMI (2016)
40. Yu, S., Tan, D., Tan, T.: A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In: ICPR (2006)
41. Zhang, S., Wang, Y., Li, A.: Cross-view gait recognition with deep universal linear embeddings. In: CVPR (2021)
42. Zhang, Y., Huang, Y., Wang, L., Yu, S.: A comprehensive study on gait biometrics using a joint cnn-based method. PR (2019)
43. Zhang, Z., Tran, L., Liu, F., Liu, X.: On learning disentangled representations for gait recognition. TPAMI (2020)
44. Zhang, Z., Tran, L., Yin, X., Atoum, Y., Liu, X., Wan, J., Wang, N.: Gait recognition via disentangled representation learning. In: CVPR (2019)
45. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: AAAI (2020)
46. Zhou, K., Yang, Y., Cavallaro, A., Xiang, T.: Omni-scale feature learning for person re-identification. In: ICCV (2019)