Pattern Recognition 125 (2022) 108519

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

A unified perspective of classification-based loss and distance-based loss for cross-view gait recognition

Feng Han[a,b], Xuejian Li[a,b], Jian Zhao[d,*], Furao Shen[a,c,**]

[a] State Key Laboratory for Novel Software Technology, Nanjing University, China
[b] Department of Computer Science and Technology, Nanjing University, China
[c] School of Artificial Intelligence, Nanjing University, China
[d] School of Electronic Science and Engineering, Nanjing University, China

ARTICLE INFO

Article history:
Received 25 December 2020
Revised 26 October 2021
Accepted 29 December 2021
Available online 1 January 2022

Keywords:
Biometrics
Gait recognition
Computer vision
Metric learning
Angular softmax loss function
Triplet loss function

ABSTRACT

Gait can be used to recognize people in an uncooperative and noninvasive manner and it is hard to imitate or counterfeit, which makes it suitable for video surveillance. The current solutions for gait recognition are still not robust to handle the conditions when the view angles of the gallery and query are different. We improve the performance of cross-view gait recognition from the perspective of metric learning. Specifically, we propose to use angular softmax loss to impose an angular margin for extracting separable features. At the same time, we use triplet loss to make the extracted features more discriminative. Additionally, we add a batch-normalization layer after extracting gait features to effectively optimize two different losses. We evaluate our approach on two widely-used gait dataset: CASIA-B dataset and TUM GAID dataset. The experiment results show that our approach outperforms the prior state-of-the-art approaches, which shows the effectiveness of our approach.

1. Introduction

Gait, which is the way people walk, can be used to identify person's identity. Compared to other biometrics, gait recognition has several benefits. A unique advantage of gait as a biometric is that it offers the potential to recognize people at a distance or at a low resolution, as long as the video can capture the whole body movement, while other biometrics might not be perceivable. Also, gait recognition can identify human in an uncooperative and noninvasive manner. It can recognize people without the subject even knowing, while other biometrics like face or fingerprint require explicit cooperation from the subject. Furthermore, gait is a dynamic behavioral feature, which makes it hard to camouflage or imitate.

Most of the methods that solve the problem of gait recognition can be divided into the following two categories: model-based approaches and appearance-based approaches.

Model-based approaches first model human body utilizing the 2D or 3D structure of the human body, and then use the model to capture discriminative features. Model-based approaches are robust against appearance variations and view variations. However, they are usually sensitive to the effectiveness of the model and require the camera to capture the video well enough to generate the model.

On the other hand, appearance-based approaches directly extract discriminative features from the video. Due to the fact that they do not need to explicitly build the model, they require lower image resolution than model-based approaches. The most typical appearance-based method is the Gait Energy Image (GEI) [1] template. It first segments the image to get the sequence of walking silhouettes, then averages the sequence along the time dimension to get the GEI template and uses Principal Component Analysis (PCA) or Multiple Discriminant Analysis (MDA) to perform dimensionality reduction. The vector after transformation is the final representation for the gait sequence. Shiraga et al. [2] trained an eight-layer neural network to classify the GEI template. Simple as it is, it can achieve promising results on some simple experiment settings. However, due to the averaging process, GEI template will lose dynamic information of the sequence.

There are many factors that can affect people's walking appearance, such as walking speed, clothes, shoes, carrying bags and view angles. When recognizing person with a different view from the gallery, the recognition rate drops significantly [3]. In order to make gait recognition more suitable for practical applications, a cross-view gait recognition system that is robust to bag-

* Corresponding author at: School of Artificial Intelligence, Nanjing University, China.
** Corresponding author at: School of Electronic Science and Engineering, Nanjing University, China.
E-mail addresses: fenghan@smail.nju.edu.cn (F. Han), lixj@smail.nju.edu.cn (X. Li), jianzhao@nju.edu.cn (J. Zhao), frshen@nju.edu.cn (F. Shen).

https://doi.org/10.1016/j.patcog.2021.108519
0031-3203/

carrying condition variations and coat-wearing condition variations is needed.

Recently, deep neural networks have been widely used in a variety of computer vision tasks, including face recognition, image classification, image segmentation, etc. Approaches based on deep learning fall into the second category, i.e., appearance-based approaches. Due to the fact that traditional CNNs can only process image data while gait recognition often requires processing video data, how to fuse the temporal information is a problem that must be solved. Wu et al. [4] proposed 3D-CNN architectures to extract features from frames and fuse temporal information. However, this type of architecture can only take a fixed number of frames as input. The general solution is either random sampling from the sequence, or using a moving window to split the whole sequence and then average the resultant feature vectors. Due to the fixed length of input sequence, 3D CNN is usually unable to access the whole sequence which will lead to information loss. Other researchers use the LSTM (Long Short-Term Memory) to modeling temporal information [5–7]. However, LSTM architectures are not able to process temporal data in parallel, which is very time-consuming. Chao et al. [8] treated gait sequence as a set, based on the assumption that the appearance of a silhouette already contained its positional information. Under this assumption, they designed an effective architecture called GaitSet to extract feature from a set of walking silhouettes and achieved state-of-the-art results. However, they only used triplet loss [9] to guide the network to extract discriminative features. It still suffers from a weak generalization capability from the training set to the testing set.

In this paper, we propose a new loss function for gait recognition, which can guide the network to map the gait feature to a separable yet discriminative feature space. The loss function utilize angular softmax (A-Softmax) loss [10], which learns a separable feature in the cosine space, and triplet loss [9] to increase the distance between feature vectors of different subjects and decrease the distance between the feature vectors of the same subject. Our approach uses the most effective architecture of gait recognition to date, GaitSet, as our backbone network. Using our proposed loss function, our results on CASIA-B dataset [3] under cross-view scenarios with bag-carrying condition variations and coat-wearing condition variations and TUM Gait from Audio, Image and Depth (GAID) dataset [11] exceed the previous state-of-the-art results. Our main contributions can be summarized as follows:

- We propose a new loss function that guides the network to map the gait feature to a separable yet discriminative feature space. Our loss function utilizes A-Softmax to learn a separable feature in the cosine space and triplet loss to increase the distance between feature vectors of different subjects and decrease the distance between the feature vectors of the same subjects.
- A-Softmax and triplet loss are optimized in different spaces. In order to make the training process feasible, we add a batch-normalization layer after extracting gait feature (before the last fully-connected layer) to reduce the impact of optimizing two different losses.
- We conduct comprehensive experiments on CASIA-B dataset and TUM GAID dataset. CASIA-B dataset contains 11 views of each walking sequence and 3 different appearance/condition variations (i.e., normal, bag, and coat). Although the TUM GAID dataset just has one view angle, it can be used to test the generalization performance of the model. The experiment results show that using our loss function with GaitSet as our backbone network exceeds the previous state-of-the-art performance under the same experiment settings.

The rest of this paper is organized as follows. We present related works in Section 2. In Section 3, we introduce our proposed approach. The experiment results showing the effectiveness of our approach are presented in Section 4. Finally, we present conclusions and discussions in Section 5.

2. Related work

The researches on gait recognition can be roughly divided into two categories: model-based approaches and appearance-based approaches.

2.1. Model-based approaches

Model-based approaches first model the human body utilizing the 2D or 3D structure of human model, and then exploit the model to generate discriminative features.

Dockstader et al. [12] introduced the concept of soft kinematic constraints by using a hierarchical, structural model of the human body. Luo et al. [13] used multi-view gait silhouettes to reconstruct 3D parametric gait model by an optimized 3D human pose, shape and simulated clothes estimation method. Ariyanto and Nixon [14] used a structural model including articulated cylinders with 3D Degrees of Freedom (DoF) at each joint to model the human lower legs. Wang and Yan [15] proposed a novel gait representation called area average distance and ensembled the results of several view-specific Hidden Markov Model (HMM) gait learners to improve the cross-view gait recognition performance. Furthermore, Liao et al. [16] recently proposed PoseGait exploiting human 3D pose estimated from images by CNN as the feature for gait recognition.

Model-based approaches are robust against appearance variations and view variations. However, they are usually sensitive to the effectiveness of the model and require the camera to capture the video well enough to fit the model.

2.2. Appearance-based approaches

Appearance-based approaches (also known as model-free approaches) aim to directly extract discriminative features from the video. Spatial-temporal templates are the most widely-used methods. Han et al. proposed GEI [1] template, which is the average of the silhouettes along time dimension and achieved promising results under some simple experiment settings. Despite its great success, GEI template is sensitive to some variations like clothing and view angles because the averaging process can cost information loss.

To make gait recognition more robust to view variations, View Transformation Models (VTM) [17] is proposed to transform the view angle of the probe gait templates to the gallery corresponding view angle. However, the overall performance of gait recognition essentially depends on the quality of the transformed results.

With the rapid development of convolutional neural networks, more and more researchers use deep learning based approaches to handle gait recognition problems. The deep learning based approaches can be further divided into two categories:

2.2.1. Generative approaches

Generative approaches usually transform different gait representations of different view angles or different conditions to a common view angle or condition. Yu et al. proposed to use a generative adversarial networks (GAN) [18] to transform GEI template from any given view angles to the same view angle and from any given carrying conditions or clothing condition to normal walking condition. He et al. [19] proposed a multi-task generative adversarial network to transform view-specific gait features to another view, based on the assumption that gait images with view variations lie on a low-dimensional manifold. Zhang et al. [20] proposed View Transformation GAN (VT-GAN), which utilizes triplet

loss to preserve identity information. Han et al. [21] proposed a multi-view alternating back-propagation algorithm to learn multi-view generator networks by allowing them to share common latent factors. The drawback of the generative approaches is that it is sensitive to the generated images or features and sometimes are not optimal due to the generation process.

2.2.2. Discriminative approaches

Discriminative approaches aim to directly learn discriminative gait representations. Wu et al. [4] first introduced deep CNNs to gait recognition. They trained several networks to extract the similarity of two input sources (e.g., GEI templates or sampled silhouettes). Takemura et al. [22] carefully designed networks using contrastive loss for the gait verification task and triplet ranking loss for the gait recognition task. Depending on the view angle, they further designed high-level and low-level structures to explicitly choose when to compute the difference of two input features. This approach requires the knowledge of probe and gallery view angle. Castro et al. [23] used a CNN to learn high-level descriptors from stacked optical flow components. Zhang et al. [6] explicitly disentangled temporal-based gait feature and frame-based appearance feature by an auto-encoder network. Chao et al. [8] proposed to view the sequence of walking silhouettes as a set and to directly extract feature from the set, based on the assumption that the appearance of one silhouette already contained its positional information. Zhang et al. [7] proposed to learn effective spatial-temporal features by a learned horizontal partition and combine the weighted results of each part by a score learned by an LSTM attention model as the final representations of gait features. Although deep neural network based models are able to achieve good performance in representation learning, deep learned features cannot be interpreted easily. To meet this demand, Yuan et al. [24] proposed a Gabor convolution module which showed both good interpretability and superior performance.

2.3. Metric learning

Gait recognition is a metric learning problem. The purpose of metric learning is to learn a map function to transform the input to a feature space, where the distance among features can represent the similarity among the inputs. The metric learning has been applied in a wide range of fields, including gait recognition, face recognition, and person re-identification, etc. For instance, Zhang et al. [25] proposed a new cost function for metric learning used in person re-identification. The proposed method formulates it as a constrained optimization problem by imposing a constraint on the linear transformation. The proposed cost function can be solved by an efficient matrix optimization method.

Loss functions play an important role in deep metric learning as they can guide the network parameters' update direction. Choosing a proper loss function is essential for the performance. Typically, there are two types of loss functions: classification-based and distance-based.

2.3.1. Classification-based loss functions

Softmax loss function is the most popular loss function for classification problems. It is very intuitive because for each input tensor, it can compute a classification score for each class. For metric learning problems, during testing stage, the last fully-connected layer is removed and its prior layer's output is the extracted feature vector. However, using softmax loss can only learn separable features [26] (i.e., can be separated using a decision boundary). For metric learning problems, we want the features to be discriminative (i.e., the distance between features of the same subject is smaller than that of different subjects). To alleviate this problem, Liu et al. [27] introduced the Euclidean margin to softmax loss, and achieved promising result. However, the Euclidean space may not be suitable for learning discriminative features [10]. Liu et al. [10] proposed A-Softmax loss to learn angularly distributed features, and introduced the angular margin in the cosine space. After that, Wang et al. [28] changed the form of applying the margin, while the margin is still applied in the cosine space.

2.3.2. Distance-based loss functions

Distance-based loss function directly aims at learning discriminative features, which is more suitable in metric learning problems. Wen et al. [26] proposed center loss to optimize the distance between samples in the same category, by assigning a category center to each category. Center loss only considered the intra-class distance, while the inter-class distance is also an important part to meet the goal of metric learning. Schroff et al. [29] proposed triplet loss to minimize the intra-class distance and at the same time make the inter-class distance larger than the intra-class distance by a margin. In order to ensure fast convergence, it is crucial to select triplets that violate the triplet constraint [29]. Triplet loss is difficult to optimize and mining hard triplets is crucial for learning. Hermans et al. [9] introduced variants of the classic triplet loss which makes mining of hard triplets unnecessary. They used only a triplet loss and no special layers to achieve state-of-the-art results for the human re-identification task by selecting the hardest positive and the hardest negative samples within a mini-batch when forming a triplet for computing the loss. He et al. [30] proposed the triplet-center loss (TCL) to learn a center for each class and aims at making the distances between an instance and its corresponding center (instead of a positive instance) smaller than the distance between it and the center of different classes (instead of a negative instance). TCL can avoid the complex construction of triplets and the necessity of mining hard samples. Furthermore, Li et al. [31] used the cosine distance instead of the Euclidean distance to compute the distance between instance and class centers, and they achieved better results. Zhang et al. [7] proposed angle center loss (ACL) for cross-view gait recognition problems. It assigns a center to each view angle of each subject, and minimizes the maximum distance between features of the same subject's different view angles. However, they only considered a cross-view settings while the bag-carrying condition and clothing condition are also essential to gait recognition problems.

3. Proposed approach

3.1. Framework

Our approach uses GaitSet [8] as the backbone network. The GaitSet treats silhouette sequence as a set and is currently the most effective network for gait recognition. Based on this network architecture, we propose a new loss function. The framework of our approach is shown in Fig. 1. Our overall framework is the same as the original GaitSet, the input of the network is a sequence of walking silhouettes, and the output is a set of features in different scale. During the testing stage, the features are concatenated into the final representation to compute the distance.

3.2. Review of loss functions

3.2.1. Softmax loss

Softmax loss first normalizes the network output to $[0, 1]$, which can be viewed as the probability, then computes the logarithmic likelihood based on the probability. Let K denote the number of subjects in the training set, N denote the number of samples

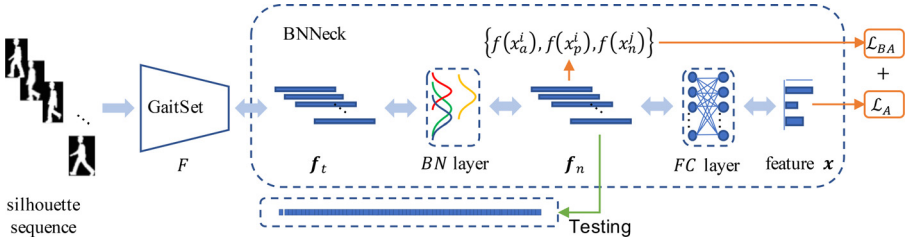

Fig. 1. The framework of our proposed approach. Our loss function contains two parts: triplet loss and angular softmax loss. Due to the fact that triplet loss is optimized in the Euclidean space and angular softmax loss is optimized in the cosine space, we add a batch-normalization layer before the last fully-connect (FC) layer to make the training process feasible.

in the training set. Then softmax loss can be calculated as (1).

$$\mathcal{L}_{softmax} = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i + b_{y_i}}}{\sum_{j=1}^{K} e^{\boldsymbol{W}_{y_j}^T \boldsymbol{x}_i + b_{y_j}}}\right), \tag{1}$$

where $\boldsymbol{W}$ denotes the parameters of the last fully-connected layer; $\boldsymbol{x}_i$, $i \in \{1, \ldots, N\}$, represents the extracted feature vector for the ith walking sequence in the training set; y_i denotes the corresponding label of feature $\boldsymbol{x}_i$; $\boldsymbol{W}_j$, $j \in \{1, \ldots, K\}$, denotes the jth column of the weight parameters $\boldsymbol{W}$; $\boldsymbol{b}_j$, $j \in \{1, \ldots, K\}$, denotes the jth element of the bias of the last fully-connected layer.

The deep features learned by softmax loss are separable for classification tasks, but it is hard to measure the similarity of them. In other words, softmax has not considered the distances among intra-class samples, so it cannot achieve intra-class compactness and inter-class separability.

3.2.2. Modified softmax loss

The original softmax loss can only learn separable features in the Euclidean space, which may not be an optimal solution. The last fully-connected layer can be viewed as computing the cosine distance of the corresponding class [32]. Performing optimization in the cosine space might be a better option. To achieve this goal, just normalize $\left\|\boldsymbol{W}_j^T\right\|_2$ to 1, $\forall j \in \{1, \ldots, K\}$ in each iteration and set the biases to zero. The modified softmax loss can be computed as

$$\begin{aligned}\mathcal{L}_{modified-softmax} &= \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\boldsymbol{W}_{y_i}^T \boldsymbol{x}_i}}{\sum_{j=1}^{K} e^{\boldsymbol{W}_{y_j}^T \boldsymbol{x}_i}}\right) \\ &= \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\|\boldsymbol{x}_i\| \cos(\theta_{y_i,i})}}{\sum_{j=1}^{K} e^{\|\boldsymbol{x}_i\| \cos(\theta_{j,i})}}\right),\end{aligned} \tag{2}$$

where $\cos(\theta_{j,i})$ denotes the angle between the extracted feature $\boldsymbol{x_i}$ and the jth column of the fully-connect layer parameters $\boldsymbol{W}$.

Comparing with the original softmax, the modified softmax changes from optimizing inner product to optimizing angles. The features learned by the modified softmax are angular distributions, but may not be discriminative enough.

3.2.3. Angular softmax loss (A-Softmax)

Angular softmax loss was first introduced in [10] to make the features further separable (i.e. maximizing inter-class distance and minimizing intra-class distance) by imposing an angular margin. The angular softmax loss can be calculated as (3).

$$\mathcal{L}_A = \frac{1}{N}\sum_{i=1}^{N} -\log\left(\frac{e^{\|\boldsymbol{x}_i\| \psi(\theta_{y_i,i})}}{e^{\|\boldsymbol{x}_i\| \psi(\theta_{y_i,i})} + \sum_{j \neq y_i} e^{\|\boldsymbol{x}_i\| \cos(\theta_{j,i})}}\right), \tag{3}$$

where $\psi(\theta_{y_i,i}) = (-1)^k \cos(m\theta_{y_i,i}) - 2k$, $\theta \in [\frac{k\pi}{m}, \frac{(k+1)\pi}{m}]$ and $k \in \{0, \ldots, m-1\}$, where $m \geq 1$ is an integer hyperparameter that controls the size of the angular margin.

The A-Softmax utilizes an angular margin between the ground truth class and other classes to maximize the distances among inter-class samples. It makes the decision boundary more separated and stringent. However, A-Softmax is usually unstable and the setting of hyperparameter m needs to be carefully adjusted. It uses a fixed value to enlarge the margin for all classes, which may not be suitable for all situation in practice. Moreover, it only uses the ground truth class to enlarge the feature margin, which does not make the best use of the non-ground truth classes.

3.2.4. Contrastive loss

Contrastive loss [33] is a distance-based loss function and is intended for siamese networks, whose input is a pair of samples. The groundtruth relationship y equals 1 if the two samples are from the same class and 0 otherwise. The contrastive loss can be computed using (4). It tries to directly minimize the distance between two samples belonging to the same class, and make the distance between two samples from different classes larger than a given margin m.

$$\mathcal{L}_c = yd^2 + (1-y)\max(m-d, 0)^2, \tag{4}$$

where d is the Euclidean distance between two samples.

Compared with the classical classification loss functions, contrastive loss directly minimizes the distances among intra-class samples and maximizes the distances among inter-class samples. As a result, it is able to map similar samples to nearby points on the output manifold and dissimilar samples to distant points. However, the contrastive loss uniformly makes the distance of all the same class samples to 0 and the different class to a fixed margin. This is a strict constraint for some samples having a larger difference with other samples of the same class.

3.2.5. Batch-All (BA) triplet loss

Triplet loss is another widely-used loss function for metric learning. Given an anchor sample, a positive sample from the same class as the anchor and a negative sample from a different class, triplet loss tries to make the distance between the anchor and the negative sample larger than the positive sample. But the original triplet loss suffers from high computational cost. The performance is very sensitive to the sample mining strategies which require careful design. To avoid this problem, Hermans et al. [9] proposed the Batch-All (BA) triplet loss, which utilizes online hard example mining for triplet loss on entire mini-batch. It first samples a mini-batch which contains P people and each person has Q walking sequences, then computes the total triplet loss with every possible sample treated as an anchor. The BA triplet loss can be computed as follows:

$$\mathcal{L}_{BA} = \overbrace{\sum_{i=1}^{P}\sum_{a=1}^{Q}}^{\text{all anchors}} \overbrace{\sum_{\substack{p=1 \\ p\neq a}}^{Q}}^{\text{all pos.}} \overbrace{\sum_{\substack{j=1 \\ j\neq i}}^{P}\sum_{n=1}^{Q}}^{\text{all negatives}} \left[m - d_{j,a,n}^{i,a,p}\right]_+, \tag{5}$$

$$d_{j,a,n}^{i,a,p} = D(f(x_a^i), f(x_p^i)) - D(f(x_a^i), f(x_n^j)), \tag{6}$$

where $D(\cdot,\cdot)$ denotes the Euclidean distance between two samples; x^i is the sampled input data of person i; $f(\cdot)$ is the neural network that extracts features; m is a margin controls how much the distance between the anchor and the positive sample is greater than the distance between the anchor and the negative sample we want.

3.3. Proposed loss function

Combining the advantages of the above loss functions, we can guide the network to learn a separable and at the same time discriminative feature representation. More specifically, by utilizing A-Softmax loss, our network can learn a separable feature representation in the angular space, and by utilizing BA triple loss, we can make our learned feature representations more discriminative. The final loss function can be expressed as follows:

$$\mathcal{L} = \mathcal{L}_{BA} + \alpha \mathcal{L}_A, \tag{7}$$

where α is a hyperparameter that controls the ratio of Batch-All triplet loss function and the angular softmax loss function.

3.4. Batch normalization neck

Angular softmax loss mainly optimizes the distance in the cosine space while triplet loss focuses optimizing the distance in the Euclidean space. Inspired by Luo et al. [34], in order to make the training process feasible, we added a batch normalization neck (BNNeck) after extracting gait features.

Batch normalization first normalizes the output of the previous layer by subtracting the mini-batch mean and dividing the mini-batch standard deviation, then uses two trainable parameters to control the mean and standard deviation of the new mini-batch data, respectively. It can reduce the internal covariate shift and thus accelerate deep network training. It has a great feature: the output changes of the previous layers of the network will not have much impact on the latter layer. In our case, although two loss functions are optimized in different spaces, batch-normalization layer can reduce the impact of optimizing two different losses, and thus make training process feasible.

We add a batch-normalization layer after extracting gait feature (which is denoted as $\boldsymbol{f}_t$). The resulting normed gait feature is denoted as $\boldsymbol{f}_n$. Both $\boldsymbol{f}_t$ and $\boldsymbol{f}_n$ can be used for identifying human.

3.5. The system

Like other recognition systems, our system contains two stages: training stage and testing stage. Next we will introduce these two stages in detail.

3.5.1. Training stage

The training process is divided into two main steps: alignment and network training. More specifically, in order to remove the effect of the distance between the camera and the subject, all the silhouettes are aligned to a uniform size of 64×44 based on the methods in [35]. Then, we randomly sample P people and sample Q silhouettes sequences from each person. After that, the sampled silhouettes sequences are combined into a tensor and passed to the network to extract features and classification vectors. In particular, in each iteration, we first use the features to compute the BA triplet loss with (5). Then, we use the classification vectors to compute the angular softmax loss with (3). The joint optimization process is conducted until converged. For the detailed training process, see Algorithm 1.

Algorithm 1 Training Process.

Input: Training data D, which contains N walking silhouettes sequences of K people, batch size (P, Q), backbone network F, a BN layer BN, a FC layer FC, how many training epochs e, the ratio of different loss functions α.
Output: Trained network $F_\Theta, BN_\Theta, FC_\Theta$.

```
for i in [1, N] do
    D̃_i ← align(D_i)
end for
for i in [1, e] do
    Random sample P people from N;
    for j in [1, P] do
        Random sample Q silhouettes sequences from person P_j;
    end for
    Combine sampled silhouette sequences into a tensor T;
    Pass T to the network, f_t ← F(T);
    Input f_t to the BN layer, f_n ← BN(f_t);
    Input f_n to the FC layer, x ← FC(f_n);
    Use f_n to compute the BA triplet loss L_BA with (??);
    Use x to compute the angular softmax loss L_A with (??);
    Total loss L ← L_BA + αL_A
    Update network parameters Θ with backward propagation;
end for
return the trained network F_Θ, BN_Θ, FC_Θ.
```

3.5.2. Testing stage

The testing process is divided into three steps: alignment, feature extraction and search in the gallery. We first align the silhouettes to a fixed size same as in the training stage. Then, the probe walking silhouettes are passed to the trained network to extract gait feature. In the end, search the gallery for the nearest neighbor of the probe feature. The distance between the probe feature and its nearest neighbor is compared with a predetermined threshold for verification. If the distance is smaller than the threshold, the result of recognition is the identity of its nearest neighbor. For the detailed testing process, see Algorithm 2.

Algorithm 2 Testing Process.

Input: Trained network $F_\Theta, BN_\Theta, FC_\Theta$, gallery walking silhouettes sequences G, which contains N sequences, label of all gallery videos L, query walking sequence q, distance metric $D(\cdot,\cdot)$, a distance threshold t.
Output: If the query person is in the gallery, return the label, else return None.

```
q̃ ← align(q);
Feature f_{n_q} ← BN_Θ(F_Θ(q̃));
for i in [1, N] do
    G̃_i ← align(G_i);
    Feature f_{n_i} ← BN_Θ(F_Θ(G̃_i));
    Distance d_i = D(f_{n_q}, f_{n_i});
end for
m ← argmin_i d;
if d_m ≤ t then
    return L_m.
else
    return None.
end if
```

4. Experiments

In this section, we first describe the datasets and experiment details. Then we evaluate the impact of loss functions and BNNeck by cross-view and different walking conditions experiment

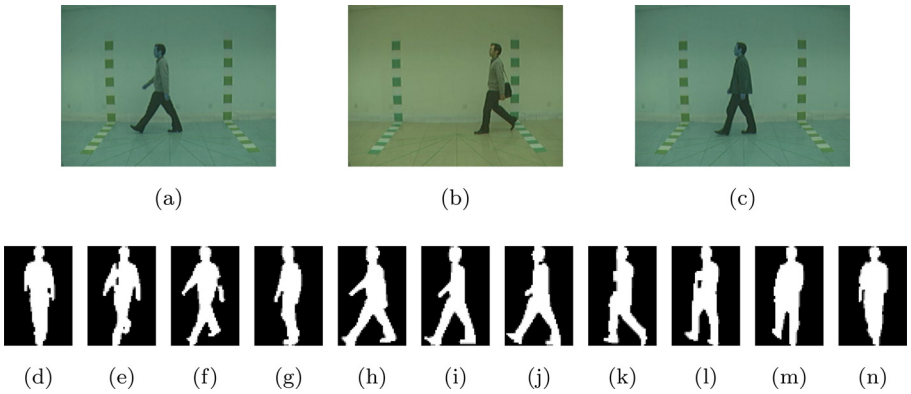

Fig. 2. Examples from the CASIA-B dataset. All these images belong to the same subject. The first row shows frames from normal walking (NM), walking carrying a bag (BG) and walking wearing a coat (CL), respectively. The second row shows aligned normal walking silhouettes captured from different view angles. The silhouettes are captured of a completed period of a subject every 18° from 0° to 180°.

settings. Finally, we compare the recognition accuracy of our proposed method with previous state-of-the-art methods.

4.1. Datasets

We conduct experiments on two widely-used large gait datasets: CASIA-B [3] database and TUM GAID [11] database. In addition, there are some other similar and excellent datasets available, such as USF HumanID Gait Challenge Dataset [36], etc.

4.1.1. CASIA-B Dataset

CASIA-B dataset is introduced in [3]. There are walking videos of 124 subjects captured from 11 view angles. In the database, three most important factors for gait recognition, i.e., view angle, clothing variations and carrying condition variations, are separately considered. For each subject, there are six normal walking videos (NM), two videos walking with a coat (CL) and two videos walking with a bag (BG), each captured from 11 views. This dataset allows us to well evaluate the impact of view, clothing and carrying condition variations.

Some examples of CASIA-B dataset are shown in Fig. 2, which shows that the view angle variations impose a great change to the appearance.

4.1.2. TUM GAID Database

TUM Gait from Audio, Image and Depth (GAID) database considers carrying condition variations and shoe type variations. It contains walking videos captured from 305 subjects and for each subject six normal walking videos (n1-n6), two walking videos carrying a backpack of approximately 5 kg (b1-b2) and two walking videos with coating shoes (s1-s2) are recorded. Although it does not involve variations in view angles, the large number of population can verify the effectiveness of our proposed approach. Examples of TUM GAID database are shown in Fig. 3.

For simplicity, we use the cropped depth images provided by the database provider and set a threshold to distinguish the subject and the background to get the silhouette.

Fig. 3. Examples from TUM GAID database. All these images belong to the same subject. The first row shows normal (N) walking style of the subject. The second row shows the subject walking carrying a backpack (B). The third row shows the subject walking wearing coating shoes (S). The first column shows one frame of a walking video stream. The second column shows the corresponding cropped depth frame of the first column provided by the database provider [11]. The third column shows the extracted silhouette.

4.2. Experiment settings

For CASIA-B dataset, same as [4,7,8], the first 74 subjects are used to train the model and the rest 50 subjects are left for evaluating the performance. During the testing stage, NM01-04 of the 50 subjects are kept as the gallery set. There are three probe sets: NM05-06 (NM), BG01-02 (BG) and CL01-02 (CL). Note that the proposed model is evaluated with cross-view scenarios, which means during the testing stage, both probe sets and gallery set contain only one specified view angle, and the view angles of any two sets are different. This is a much harder situation which enables us to demonstrate the robustness of the proposed method.

For TUM GAID dataset, we use the split configuration mentioned in [11], from which 150 subjects are used as the development set and the rest 155 subjects are used as the test set. During the testing stage, N1-N4 of the 155 subjects are kept as the gallery set. There are also three probe sets: N5-N6 (N), B1-B2 (B), S1-S2 (S).

We use GaitSet as our backbone network, and train the network using the proposed loss function. Backbone network details are the same as those in [8]. Adam is chosen as an optimizer to minimize the loss over the training data. For the hyperparameters of the Adam optimizer, the learning rate is set to be 0.0001. The batch size P and Q mentioned in Section 3 is set to be 16 and

Table 1
Averaged recognition accuracies on CASIA-B dataset using different loss functions. Results are Rank-1 recognition accuracies averaged on all 11 views. Identical view cases are excluded.

Feature	Distance Metric	NM	BG	CL
BA triplet (GaitSet)	Euclidean Distance	95.5	88.5	70.0
BA triplet (GaitSet)	Cosine Distance	94.5	88.9	68.0
A-Softmax	Euclidean Distance	92.1	86.1	59.0
A-Softmax	Cosine Distance	94.6	87.7	62.5
BA-triplet + A-Softmax	Euclidean Distance	95.8	90.6	74.2
BA-triplet + A-Softmax	Cosine Distance	**96.0**	**91.6**	**74.8**

$Q = 6$. We train the network for 80K iterations and 40K iterations for CASIA-B dataset and TUM GAID dataset, respectively.

The hyperparameters in our experiments are set with carefully studies and fine-tunings. More details about the impact of these hyperparameters can be seen in Section 4.7. In the end, the hyperparameter m that controls the size of the angular margin in A-Softmax is set to be 3. Another hyperparameter m in $\mathcal{L}_{BA}$ is set to be 0.1. The hyperparameter α that controls the ratio of different loss functions is set as 1.0 empirically.

The evaluation indicator used in the following is rank-1 recognition accuracy. Specially, results on CASIA-B dataset are averaged on the 11 gallery views excluding identical-view cases. For example, the accuracy of probe view 18° is averaged on the other 10 gallery views, excluding gallery view 18°.

4.3. Impact of loss function

We evaluate the impact of different loss functions as an ablation study. As illustrated in Table 1, using only BA triplet loss produces very promising results. This is due to the effectiveness of GaitSet architecture. As expected, using angular softmax alone can only generate separable features, not discriminative features, so the performance is much worse than using BA triplet loss. Although using only BA triplet loss may obtain discriminative features, they are still not robust enough to generalize to coat-wearing condition variations well. By combining the above two loss functions, the proposed model can learn feature vectors that are separable and at the same time discriminative. Using both loss functions achieves the best results.

We also evaluate the use of different distance metrics for computing the distance between features for identification. The Euclidean distance performs slightly better than the cosine distance for BA triplet loss, because the triplet loss is computed using the Euclidean distance. However, combining the angular softmax loss function imposes an angular margin in the cosine space, and therefore the cosine distance performs better than the Euclidean distance.

4.4. Impact of BNNeck

The output feature vector of backbone $\boldsymbol{f}_t$, is passed to a batch-normalization layer to produce another feature vector $\boldsymbol{f}_n$. During the training stage, one of $\boldsymbol{f}_t$ and $\boldsymbol{f}_n$ is used to compute the triplet loss; during the testing stage the same feature vector as the training stage is used to identify human. We evaluate the performance of $\boldsymbol{f}_t$, $\boldsymbol{f}_n$ and the feature vector without batch-normalization to demonstrate the effectiveness of the BNNeck architecture.

From Table 2, it can be seen that without batch-normalization layer, two losses cannot be effectively optimized, producing bad results. By adding a batch normalization layer, the impact of optimizing one loss on another is reduced.

We also evaluate the impact of different distance metrics. Using feature vector $\boldsymbol{f}_t$, which is the feature vector before batch normalization, as the extracted feature to compute triplet loss performs better in the Euclidean space. However, using normed feature vector $\boldsymbol{f}_n$ as the extracted feature vector achieves better results in the cosine space, which meets our expectations because of the angular margin imposed by A-Softmax. In the rest of this article, the $\boldsymbol{f}_n$ is used as the extracted feature vector.

Table 2
Averaged recognition accuracies on CASIA-B dataset using different vectors as the extracted feature vector. Results are Rank-1 recognition accuracies averaged on all 11 views. Identical view cases are excluded.

Feature	Distance Metric	NM	BG	CL
Without Batch Norm	Euclidean Distance	95.1	88.5	71.0
Without Batch Norm	Cosine Distance	95.4	88.9	72.1
$\boldsymbol{f}_t$	Euclidean Distance	95.2	89.3	71.7
$\boldsymbol{f}_t$	Cosine Distance	94.6	88.6	66.4
$\boldsymbol{f}_n$	Euclidean Distance	95.8	90.6	74.2
$\boldsymbol{f}_n$	Cosine Distance	**96.0**	**91.6**	**74.8**

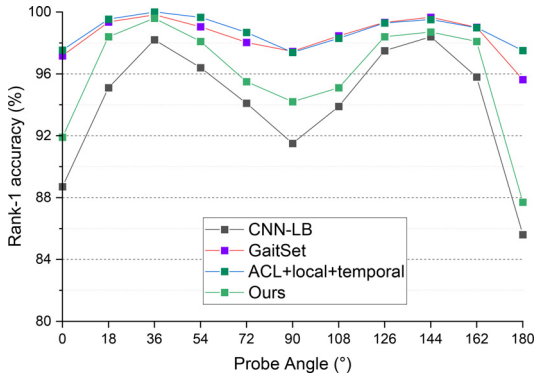

Fig. 4. Cross-view evaluations under normal walking condition in CASIA-B dataset. Gallery set: NM01-NM04, Probe set: NM05-NM06. Identical views are excluded.

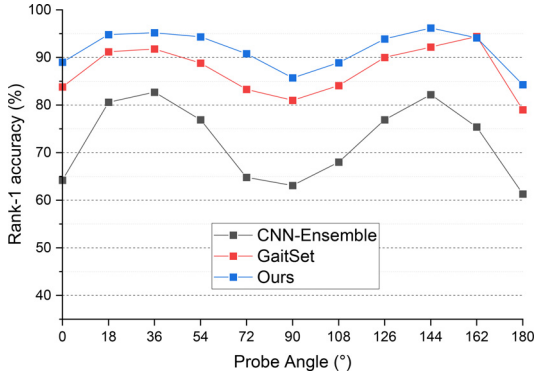

Fig. 5. Cross-view evaluations under bag-carrying walking condition in CASIA-B dataset. Gallery set: NM01-NM04, Probe set: BG01-BG02. Identical views are excluded.

4.5. Evaluations of different walking conditions

Cross-view evaluations for NM, BG, CL are shown in Figs. 4, 5 and 6, respectively. Note that each of those values is computed by averaging the accuracy among a given probe view angle with all possible different view angles of the gallery. The following conclusions can be drawn based on these results.

For normal walking conditions, our proposed approach achieves very promising results. Almost perfect accuracy is achieved under some view angles, e.g., 36° and 54°. Although the overall accuracy is the same as the previous state-of-the-art result (ACL + local + temporal), compared to their approach considering only the normal walking condition, our method additionally consider the coat-wearing condition variations and bag-carrying condition variations.

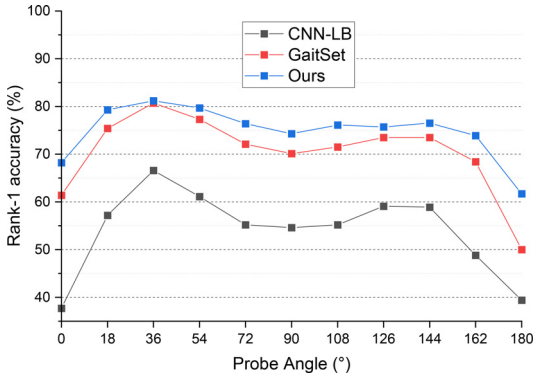

Fig. 6. Cross-view evaluations under coat-wearing walking condition in CASIA-B dataset. Gallery set: NM01-NM04, Probe set: CL01-CL02. Identical views are excluded.

It is harder to correctly recognize a person when the walking sequence of the query video and the gallery video are in different walking conditions. Specifically, the bag carried by the subject can cause a great increase of the intra-class distance. For coat-wearing conditions, the recognition accuracy drops significantly compared to normal walking condition results. This is because the long coat may greatly block the leg information, which is sometimes critical to recognize a person. There are still a lot of work to be done for gait recognition to be used in practice.

The performance of different view angles are also different. There are some view angles performing not very well, e.g., 0°, 90° and 180°. This has also been discussed in [4]. 0° and 180° contains too little useful information and profile view (90°) is the most visually different one from other views besides 0° and 180°.

The results of our proposed approach surpass the previous state-of-the-art methods in all view angles under BG and CL experiment scenarios, demonstrating the effectiveness and robustness of our approach.

4.6. Compared with state-of-the-art approaches

We compare our approach with state-of-the-art approaches on CASIA-B dataset and TUM GAID dataset.

For CASIA-B dataset, the averaged cross-view accuracy is reported. Table 3 shows that, our proposed method achieve the same accuracy with [7] for normal walking condition under cross-view settings. However, as discussed before, they only considered normal walking condition, which is not robust enough for practical gait recognition. Our approach outperforms the previous state-of-the-art approaches by 4.4 percent under bag-carrying condition variations and coat-wearing condition variations. The experiment results show that our approach is robust against view-angle variations, bag-carrying condition variations and coat-wearing condition variations.

Table 3
State-of-the-art on CASIA-B. Results are Rank-1 recognition accuracies averaged on all 11 views. Identical view cases are excluded.

Method	NM	BG	CL	Mean
MGAN [19]	68.1	54.7	31.5	51.4
CNN-Ensemble [4]	94.1	-	-	-
CNN-LB [4]	-	72.4	54.0	-
GaitSet [8]	95.0	87.2	70.4	84.2
ACL+local+temporal [7]	**96.0**	-	-	-
Ours	**96.0**	**91.6**	**74.8**	**87.5**

Table 4
Rank-1 recognition accuracy on TUM GAID database compared with State-of-the-art.

Method	N	B	S	Mean
GEI [11]	99.4	27.1	52.6	59.7
Fusion Baseline [11]	99.4	59.4	94.5	84.4
TGLSTM [5]	-	-	-	98.4
2D-CNN [37]	99.4	97.7	96.1	97.7
3D-CNN [37]	98.7	91.1	94.5	96.7
PFM [38]	99.7	99.0	99.0	99.2
CNN-SVM [23]	99.7	97.1	97.1	98.0
CNN-NN128 [23]	99.7	98.1	95.8	97.9
Ours	**100.0**	**100.0**	**99.7**	**99.9**

For TUM GAID database, Table 4 shows that our approach performs almost perfectly. For probe set N and B, our approach can correctly identify every query person, and for probe set S, our approach also performs better than the prior state-of-the-art. Note that our approach uses the provided depth images to extract walking silhouettes, which is very efficient (only need to compare the depth image with a threshold). While approaches in [37], [23] and [38] all involved computing optical flow, which is very time-consuming.

4.7. Impact of the hyperparameters

We do more experiments to analyze the impact of three hyperparameters, i.e., the angular margin factor m in (3) (it also denoted as m_1), the distance margin factor m in (5) (it also denoted as m_2) and the ratio factor α in (7) on CASIA-B dataset. The parameter m_1 varies from 1 to 5 (with 1 as the interval) and m_2, α vary from 0.1 to 1 (with 0.1 as the interval). Figure 7 show the Rank-1 recognition accuracy of different condition variations (i.e. NM, BG and CL) and the average of them with different values of these parameters. The results show that the modifications of the performance is slight when the three parameter values vary. In other words, the performance of our proposed method is relatively robust. It can

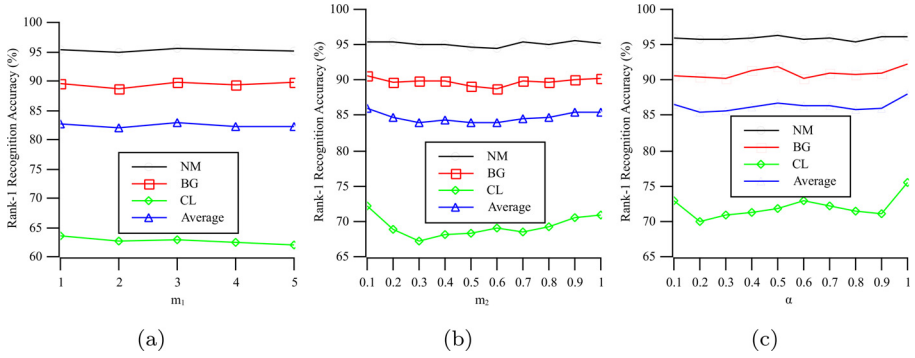

Fig. 7. Impact of three hyperparameters.

Table 5
Efficiency comparison on CASIA-B. Part of the results are reported by [7] with 4 TITAN GPUs.

Methods	Training time (h)	Testing time (min)	Feature dimension
Wu et al. [4]	7.5	40	29k
Zhang et al. [39]	0.6	3	512
Ours	3.6	0.4	256

be observed that the influences of NM and BG condition is small while the performance of CL is affected relatively heavily. This is because walking with a coat affects the appearance much more, which means it causes a more severe intra-class similarity problem than the inter-class similarity problem.

4.8. Efficiency

In the above parts, we have compared our proposed method with some state-of-the-art works in terms of accuracy. In the following, we do more experiments to explore the efficiency in terms of training time, testing time and feature dimension.

We compare the efficiency with Wu et al. [4] and Zhang et al. [39] on CASIA-B cross-view experiments. For our method, the training time is evaluated on 4 NVIDIA GeForce RTX 2080Ti GPU cards, and the testing time is conducted on a single 2080Ti GPU. On the other hand, the other two methods are implemented with more powerful Titan X GPU cards. Table 5 lists the efficiency comparison results.

As can be seen from the table, compared with the pair-wise similarity learning method [4], the proposed method is efficient in all the three aspects. It is because that our method takes advantage of a unified formula for two elemental metric learning loss functions and converges quickly. In testing stage, the proposed method only need to extract features once, and then only uses these features to compare the similarities. This saves a lot of time since it reduces the duplicated computation costs. Although our method has a slight increase of training time compared with Zhang et al. [39], the latter does not include the time of feature extraction. Moreover, there are also different experimental devices and training iterations. Our method is trained with 80K iterations, it is able to achieve comparable performances with 20K iterations, which means the training time will decrease to 0.9 h. However, the proposed method extracts more compact features with a dimension of 256, which is much smaller than 29k in Wu et al. [4] and 512 in Zhang et al. [39]. Therefore, our method has more superiority in terms of testing time and storage, which means it is also sufficient for real-time applications.

5. Conclusion

In this paper, we propose to make gait recognition more robust from the perspective of metric learning. The proposed method learns a mapping from silhouette sequences to discriminative embedding features based on GaitSet. The proposed model is trained with A-Softmax loss and triplet loss simultaneously. The A-Softmax loss imposes an angular margin to extract separable features, and the triplet loss reduces intra-class distance and enlarges inter-class distance to capture discriminative features. In order to make the training process feasible, we add a batch-normalization layer after extracting features. Compared with other methods, our method also achieves comparable result. Experiments on CASIA-B dataset and TUM GAID dataset show that our approach outperforms the previous state-of-the-art approaches under view-angle variations, bag-carrying condition variations and coat-wearing condition variations. The improvement of recognition accuracy demonstrates the effectiveness of our proposed method.

As far as applications, we argue that the proposed combination of classification-based and distance-based loss function can be applied to a variety of deep feature learning tasks. Especially for some tasks that learn both with class-level labels and pair-wise labels, such as person re-identification, face recognition, and fine-grained image retrieval.

In the future, we will investigate a more unified loss function of classification-based loss and distance-based loss for learning with class-level labels and pair-wise labels. In addition, we will explore more effective measures to improve the performance on some view angles, e.g., 0° and 180°.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Key Research and Development Program of China under Grant 2021ZD0201303 and National Natural Science Foundation of China under Grant 61876076.

References

[1] J. Han, B. Bhanu, Individual recognition using gait energy image, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2005) 316–322.
[2] K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, GEINet: view-invariant gait recognition using a convolutional neural network, in: International Conference on Biometrics, 2016, pp. 1–8.
[3] S. Yu, D. Tan, T. Tan, A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition, in: International Conference on Pattern Recognition, vol. 4, 2006, pp. 441–444.
[4] Z. Wu, Y. Huang, L. Wang, X. Wang, T. Tan, A comprehensive study on cross-view gait based human identification with deep CNNs, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2016) 209–226.
[5] F. Battistone, A. Petrosino, TGLSTM: a time based graph deep learning approach to gait recognition, Pattern Recognit. Lett. 126 (2019) 132–138.
[6] Z. Zhang, L. Tran, X. Yin, Y. Atoum, X. Liu, J. Wan, N. Wang, Gait recognition via disentangled representation learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4710–4719.
[7] Y. Zhang, Y. Huang, S. Yu, L. Wang, Cross-view gait recognition by discriminative feature learning, IEEE Trans. Image Process. 29 (2019) 1001–1015.
[8] H. Chao, Y. He, J. Zhang, J. Feng, GaitSet: regarding gait as a set for cross-view gait recognition, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8126–8133.
[9] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv:1703.07737(2017).
[10] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, L. Song, SphereFace: deep hypersphere embedding for face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 212–220.
[11] M. Hofmann, J. Geiger, S. Bachmann, B. Schuller, G. Rigoll, The TUM gait from audio, image and depth (GAID) database: multimodal recognition of subjects and traits, J. Vis. Commun. Image Represent. 25 (1) (2014) 195–206.
[12] S.L. Dockstader, M.J. Berg, A.M. Tekalp, Stochastic kinematic modeling and feature extraction for gait analysis, IEEE Trans. Image Process. 12 (8) (2003) 962–976.
[13] J. Luo, J. Tang, T. Tjahjadi, X. Xiao, Robust arbitrary view gait recognition based on parametric 3D human body reconstruction and virtual posture synthesis, Pattern Recognit. 60 (2016) 361–377.
[14] G. Ariyanto, M.S. Nixon, Model-based 3D gait biometrics, in: International Joint Conference on Biometrics, 2011, pp. 1–7.
[15] X. Wang, W.Q. Yan, Cross-view gait recognition through ensemble learning, Neural Comput. Appl. (2019) 1–13.

[16] R. Liao, S. Yu, W. An, Y. Huang, A model-based gait recognition method with body pose and human prior knowledge, Pattern Recognit. 98 (2020).
[17] D. Muramatsu, Y. Makihara, Y. Yagi, View transformation model incorporating quality measures for cross-view gait recognition, IEEE Trans. Cybern. 46 (7) (2015) 1602–1615.
[18] S. Yu, H. Chen, G. Reyes, B. Edel, N. Poh, GaitGAN: invariant gait feature extraction using generative adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 30–37.
[19] Y. He, J. Zhang, H. Shan, L. Wang, Multi-task GANs for view-specific feature learning in gait recognition, IEEE Trans. Inf. Forensics Secur. 14 (1) (2018) 102–113.
[20] P. Zhang, Q. Wu, J. Xu, VT-GAN: view transformation GAN for gait recognition across views, in: International Joint Conference on Neural Networks, 2019, pp. 1–8.
[21] T. Han, X. Xing, Y.N. Wu, Learning multi-view generator network for shared representation, in: International Conference on Pattern Recognition, 2018, pp. 2062–2068.
[22] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, On input/output architectures for convolutional neural network-based cross-view gait recognition, IEEE Trans. Circuits Syst. Video Technol. (2017).
[23] F.M. Castro, M.J. Marín-Jiménez, N. Guil, N.P. De La Blanca, Automatic learning of gait signatures for people identification, in: International Work-Conference on Artificial Neural Networks, 2017, pp. 257–270.
[24] Y. Yuan, J. Zhang, Q. Wang, Deep Gabor convolution network for person re-identification, Neurocomputing 378 (2020) 387–398.
[25] J. Zhang, Q. Wang, Y. Yuan, Metric learning by simultaneously learning linear transformation matrix and weight matrix for person re-identification, IET Comput. Vision 13 (4) (2019) 428–434.
[26] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, 2016, pp. 499–515.
[27] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, in: International Conference on Machine Learning, vol. 48, 2016, pp. 507–516.
[28] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, W. Liu, CosFace: large margin cosine loss for deep face recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5265–5274.
[29] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: a unified embedding for face recognition and clustering, in: IEEE conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.
[30] X. He, Y. Zhou, Z. Zhou, S. Bai, X. Bai, Triplet-center loss for multi-view 3D object retrieval, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1945–1954.
[31] Z. Li, C. Xu, B. Leng, Angular triplet-center loss for multi-view 3D shape retrieval, in: AAAI Conference on Artificial Intelligence, vol. 33, 2019, pp. 8682–8689.
[32] F. Wang, X. Xiang, J. Cheng, A.L. Yuille, NormFace: L_2 hypersphere embedding for face verification, in: ACM International Conference on Multimedia, 2017, pp. 1041–1049.
[33] R. Hadsell, S. Chopra, Y. LeCun, Dimensionality reduction by learning an invariant mapping, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1735–1742.
[34] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 1487–1495.
[35] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, Y. Yagi, Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition, IPSJ Trans. Comput. Vision Appl. 10 (2018) 4.
[36] S. Sarkar, P.J. Phillips, Z. Liu, I.R. Vega, P. Grother, K.W. Bowyer, The humanID gait challenge problem: data sets, performance, and analysis, IEEE Trans. Pattern Anal. Mach. Intell. 27 (2) (2005) 162–177.
[37] F.M. Castro, M.J. Marín-Jiménez, N. Guil, S. Lopez-Tapia, N.P. de la Blanca, Evaluation of CNN architectures for gait recognition based on optical flow maps, in: International Conference of the Biometrics Special Interest Group, 2017, pp. 1–5.
[38] F.M. Castro, M.J. Marín-Jiménez, N.G. Mata, R. Muñoz-Salinas, Fisher motion descriptor for multiview gait recognition, Int. J. Pattern Recognit. Artif. Intell. 31 (01) (2017) 1756002:1–1756002:40.
[39] Y. Zhang, Y. Huang, L. Wang, S. Yu, A comprehensive study on gait biometrics using a joint CNN-based method, Pattern Recognit. 93 (2019) 228–236.

Feng Han received the B.Sc. degree from Sichuan University, Sichuan, China, in 2017, and is currently pursuing the Ph.D. degree at the Nanjing University, Nanjing, China. His current research interests include multimodal learning and computer vision.

Xuejian Li received the B.Sc. degree from Nanjing University of Posts & Telecommunications, Nanjing, China, in 2018, and the M.Sc. degree from the Nanjing University, Nanjing, China, in 2021. His current research interests include gait recognition and computer vision.

Jian Zhao (Senior Member, IEEE) received the B.S. degree from Nanjing University, Nanjing, China, the M.Sc. degree from the Hamburg University of Technology, Hamburg, Germany, and the Dr. Sc. degree in electrical engineering from the Swiss Federal Institute of Technology (ETH) Zurich, Switzerland. From 2010 to 2015, he was a Research Scientist with the Institute for Infocomm Research, A*STAR, Singapore. Currently, he is an Associate Professor with the School of Electronic Science and Engineering, Nanjing University. His research interests include deep neural networks, mathematical optimization, and wireless communication networks. Dr. Zhao was honored with the Dengfeng Scholars Program of Nanjing University in 2015, IEEE Globecom 2008 Best Paper Award, and the 2009 Chinese Government Award for Outstanding Self-Financed Students Abroad.

Furao Shen (Member, IEEE) received the B.Sc. and M.Sc. degrees in mathematics from Nanjing University, Nanjing, China, in 1995 and 1998, respectively, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 2006. He is currently a Full Professor of computer science and technology with Nanjing University. His current research interests include neural computing and robotic intelligence.