Joint Metric Learning and Hierarchical Network for Gait Recognition

HUANHUAN XU1, YUQIAN LI2, XUEMEI SUN1, AND SHENGJIN WANG3, (Senior Member, IEEE)

1School of Computer Science and Technology, Tiangong University, Tianjin 300387, China
2School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
3Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

Corresponding author: Shengjin Wang (wgsgj@tsinghua.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61701277 and Grant 61771288, in part by the Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant BNR2019TD01022, and in part by the Research Fund from the Institute for Guoqiang, Tsinghua University under Grant 2019GQG0001.

ABSTRACT Gait recognition aims to identify person by walking pattern of individuals, which has a valuable application prospect in the fields of identity identification, public security and medical diagnosis. However, this paper focuses on the feature space of gait recognition feature learning, and optimizes the feature space by introducing some general criteria to improve the discrimination of depth features. In recent literature, optimizing the feature space has been proven to be beneficial for personal identification. Taken above insights together, we propose a novel gait recognition method based on joint metric learning and hierarchical network in an end-to-end manner, for improving the ability of feature identification. The contribution of this paper is two-fold. First, a joint metric learning and hierarchical network is used to realize the learning of fine-grained information and accelerated the discriminative feature learning. Second, to learn discriminative feature, we use a fuse loss function, the triplet loss and classifier loss. It guides the training process, pulls the samples in the same set close to each other, pushes the samples in different sets away from each other, and takes the silhouettes in each set as a whole. We evaluate the proposed method on two widely-used gait datasets, ie. CASIA-B and OU-LP-Bag. Experimental results demonstrate the effectiveness of the proposed method, significantly improve gait recognition performances under cross-view and bag-carrying walking conditions, respectively. Particularly, we achieves Rank-1 accuracy of 93.2% on OU-LP-Bag dataset, better than existing state-of-the art.

INDEX TERMS Metric learning, discriminatory feature, gait recognition, cross-view, carrying.

I. INTRODUCTION

Gait [1] is a unique biological feature that can identify at a distance, non-invasive and without subject’s cooperation. Gait recognition aims to identify person by walking pattern of individuals. Therefore, it is a valuable application prospect in the fields of identity identification, public security and medical diagnosis. Nonetheless, in real-world scenarios, the accuracy of gait recognition is easily affected by external factors, such as camera viewpoints, walking speed, clothing, and carrying condition.

To alleviate these issues, latest literatures usually compress the whole gait silhouettes into one image or preprocesses them with a gait template on gait recognition extract feature [2]–[5]. These methods can be performed easily, while losing temporal and spatial information. And extracts the gait features [6]–[8] directly from the original gait silhouette sequences, such as using 3D-CNN [9], [10], which is a better way to learn the features. Previous methods [6], [7] have been used loss function to improve discriminative feature, while these methods use only one loss function, the classification effect is not obvious, to illustrate this more explicitly, compared to using triplet or cross-entropy loss alone, the combination of two loss is more efficient, as shown in Fig. 1. It is more difficult to obtain the well-trained model. These methods are mainly based on data processing and network design, and are complex and difficult to execute. In addition, these methods without effectively solve the difficulties in gait recognition, such as cross view and carrying condition.
To this end, there are few studies that focus spatial optimization of features. Therefore, a joint metric learning and hierarchical network approach is proposed for gait recognition, this method improve the ability to identify features by improving the loss function of deep metric learning. Our framework consists of two modules which are responsible for visual representation and high-level semantic metric, respectively. The visual representation module is obtained higher-level feature maps, it can retain temporal and spatial information better than base on gait template. The high-level semantic metric module is designed to help deep networks extract both local and global features. After the global pooling of visual features, the feature embedding space is learned by using the double layer full connection (FC) layer. We further propose an optimal consist of triplet loss and cross entropy loss to accelerate discriminative feature learning. It guides the training process, pulls the samples in the same set close to each other, pushes the samples in different sets away from each other, and takes the silhouettes in each set as a whole. Therefore, this architecture design optimizes the gait features to reduce the intra-class distance and increase the inter-class distance, as shown in Fig. 2. Compared to using triplet or cross-entropy loss alone, the combination of two loss is more efficient. The comparison will be shown in Sec.IV-D. The advantages of the proposed method are summarized as follows:

- To focus on the feature space of gait recognition feature learning, and improve the feature recognition ability by improving the loss function of deep metric learning. This paper propose a joint metric learning and hierarchical network for gait recognition, which can learn discriminative feature.

- To learn discriminative feature, we use a fuse loss function, the triplet loss and classifier loss. It guides the training process, pulls the samples in the same set close to each other, pushes the samples in different sets away from each other, and takes the silhouettes in each set as a whole.

- We evaluate the proposed method on CASIA-B and OU-LP-Bag which are two widely-used datasets, the proposed method can significantly improve recognition
performances under cross-view and bag-carrying walking conditions, respectively. In particular, we achieve a Rank-1 accuracy of 93.2% on OU-LP-Bag, better than existing state-of-the-art.

II. RELATED WORK

A. MAIN METHODS OF GAIT RECOGNITION

Most of the current methods for gait recognition enhance the discrimination of visual features. Zhang et al. [7] presented a method based on automatic encoding by separates appearance and posture features from RGB frames, and then combines them together to form a gait feature that is a video segment using LSTM. To extract discriminative gait features, Yu et al. [11] presented a GaitGAN method based on generation of antagonistic networks. A multitask generation antagonistic network (MGAN) [3], [12] for learning perspective representation is presented. The (template) generation aims to compress the original gait silhouettes into a single image, or a single template, such as a gait energy image (GEI) [5] and a timed gait image (CGI) [13]. In the template matching method, the Perspective Transform Model (VTM) [14] learned the projection between different views. A view-invariant discriminant projection (ViDP) [3] was proposed to mining the low-dimensional geometric structure of data and learning the discriminant projection matrix of view-invariant. Das et al. [15] based on deep learning is presented to accurately identify the affected frames and predict the missing frames to rebuild the complete gait periodic. Zhang et al. [8] proposed a JUCNet method, which combines the advantages of the combination of cross-gait representation and unique gait representation of single gait image. In addition, it proposed a quintuplet loss function which can minimize the class difference and maximize the class difference at the same time. Although these methods have achieved promising performances, only extract motion features and neglect some discriminative regions. The performance of appearance features is restricted. It ignores differences in gait under certain conditions, such as carrying a bag or wearing a coat. Compared with those works, our model can not only enhance the discriminative ability, but also achieve better performance.

B. DEEP METRIC LEARNING

In recent years, deep metric learning has become one of the most attractive research fields of machine learning. How to effectively metric the similarity between objects becomes the key to the problem. In face recognition, fingerprint recognition and other open classification tasks, the number of classes is often many and the number of samples intra-class is relatively small. In this case, deep learning-based classification methods often show some limitations, such as lack of intra-class constraints, difficulties in classifier optimization, and so on. These limitations can be solved by deep metric learning.

Contrast loss [16] is the beginning of deep metric learning, which first introduces deep neural networks into metric learning. Contrast loss only constrains the features of pairs within a class as close as possible, while pairs between classes have features as far as possible. Triplet loss [17], [18] considers the relative relationship between intra-class pairs and inter-class on the basis of contrast loss. Since the triplet loss function sampled three samples at a time, while taking into account both the intra-class distance and inter-class distance as well as their relative relationship, it did not make full use of all samples of each batch during training, so the paper [19] proposed to establish a dense pair-wise connection within one batch. Sohn [20] attributed the slow convergence of contrast loss and triplet to mining only one negative sample at a time during training and the lack of interaction with other negative samples. Therefore, they proposed multiple-class N-pair loss. Liu et al. [21] proposed a coupled cluster loss by first the center of positive samples is computed, and then a constructing triplet with sample center, positive samples and hardest negative samples is constructed. Kumar et al. [22] proposed a global loss and Wen et al. [23] proposed a center loss to learn a clustering center for each category of feature. Ustinova and Lempitsky [24] proposed a histogram loss, i.e., the probability of similarity obtained by random sampling in positive samples is lower than that obtained by random sampling in negative samples.

Using the second-inverse feature of the deep neural network, combined with the classifier trained by softmax and cross-entropy loss, at the same time, it can be applied to many application tasks based on deep metric learning [25]. The former is designed to find the best classification surface, while the latter is designed to learn feature embedding so that sample embedding in the same category is compact and sample embedding in different categories is far away. This prompts us to explore the relationship between metrical learning and classifier training. We find that the classifier obtained by combining the triplet loss with the cross entropy loss training is also suitable for many tasks based on deep measure learning, which further reduces the distance between classes, enlarges the class spacing, and improves the discriminant performance of the algorithm.

III. PROPOSED METHOD

A. OVERVIEW

In this work, we propose a gait recognition method based on joint metric learning and hierarchical network. The proposed method processing the input data is extract the gait features directly from the original gait silhouette sequences. Because gait is periodic motion, it can be represented by a periodic, that is, a silhouette sequence can be represented by a gait periodic, where the silhouette of each position is observed to have a unique appearance. Therefore, it is assumed that the appearance of the silhouettes contains location information. With this assumption, Chao et al. [6] has proved the validity of frame disorder in sequence, and our method continues to develop on this basis.

The deep convolution network can obtain better visual representation with a lot of labeled data and appropriate metric functions. The metric function is particularly important.
when given the training data. However latest literatures ignore reduce the intra-class distance, influence the learning of discriminative features, at the same time, there is without feature of robustness. Based on this, we design a new metric function, which combines cross entropy loss and triplet loss. The cross-entropy loss can guide the network to learn the features of each ID. However, gait is a process of motion, and gait may vary slightly among different people, which makes learning discriminatory features difficult. To end this, we use the existing work for reference and introduce triplet loss, pull the samples of the same ID closer in the feature space and push the samples of different IDs further in the feature space, thus improving the discriminatory ability of gait features.

So, the input of the framework is a set of gait silhouettes. First, the first level visual representation module is extract frame-level features of each silhouette independently and aggregate them as global features. Secondly, the high-level semantic metric module is to learn the extracted visual features and preserve the spatio-temporal information.

B. VISUAL REPRESENTATION
The visual representation module aims to obtain the gait feature representation of the gait sequence. The module can extract the features of each silhouette in the sequence and aggregate them as the global features of the sequence. Inspired by [26]–[30], In CNN, the receptive field determines the size of the input layer corresponding to an element in the output result of a certain layer, and in CNN hierarchical network, different layers have different receptive fields. The more layers, the larger the receptive field, the better the performance. In this situation, shallow features pay more attention to fine-grained, while deep features contain more global coarse-grained information. For this reason, we propose a multi-level CNN visual representation extraction module, details as shown in Table 1. The first level CNNs extract frame-level features of each silhouette independently, the second and third level CNNs aggregate frame-level features into independent video sequence features to obtain visual features.

Given any gait dataset, suppose that there are \( N \) identities in the dataset, each person is represented by \( P_i, i \in \{1, 2, \ldots, N\} \), one or more sequences corresponding to it can be marked as \( X_i = \{X_i^j | j = 1, 2, \ldots, n\} \), \( n \) denotes the number of frames in a sequence. In order to facilitate understanding, the task of obtaining visual features can be expressed as: \( H(C(X_i)) \), \( C \) is the convolution network to extract frame-level features of each gait silhouette independently. The H aims to aggregate a frame into global features through the hierarchical convolution network, thus forming higher-level visual features. The formula is \( V = H(Z) \), where V is the visual feature, \( Z = \{z^j | j = 1, 2, \ldots, n\} \) is the frame level feature. It’s worth noting that when a sequence is taken as an input, there is no spatial position relationship between gait features of a set, it is expressed as:

\[
H((z^j | j = 1, 2, \ldots, n)) = \max((x^{\pi(j)} | j = 1, 2, \ldots, n)),
\]

where \( \pi \) define as any permutation. The number of gait silhouettes of a person is arbitrary since in the real life scene, function \( H \) should be able to input any cardinal number sequence. In order to improve the performance of visual representation module, we use the global information to learn visual attention map of each frame feature map and refine it. Therefore, we use element-wise maximum (frame-max pooling) operation across the views.

C. HIGH-LEVEL SEMANTIC METRIC
The high-level semantic metric module aims to learn the extracted visual features and save the spatial-temporal information. In order to extract local and global features, the visual feature map is divided horizontally, and then each part is divided into maximum pooling and average pooling, and the results of the two parts are added correspondingly. In order to make different segmentation strips describe different receptive fields in different scales, and different segmentation strips describe different spatial positions in each scale, so the visual features are divided horizontally. At the same time, in order to obtain the discrimination of the appearance features, we design two fully connection (FC) layers with different constraint loss. We propose a new learning strategy that joint cross-entropy loss and triplet loss. The first FC layer is to map features to discrimination subspace. The second FC layer is for ID classification, so as to improve the expression ability of discrimination subspace.

Specifically, inspired by [31], the visual features are divided into \( V \) levels, where we added two layers of FC instead of just using \( 1 \times 1 \) convolution. Visual features have \( S \) scales. On the scale \( s \in \{1, 2, \ldots, S\} \), the visual feature map extracted by hierarchical network is divided into \( 2^{-s} \) strips and equally according to different strips on height dimension, that is, a total of \( \sum_{s=1}^{S} 2^{-s} \) strips. Specifically, assume each segmented strips as \( V_{s,t}, t \in \{1, 2, \ldots, 2^{-s}\} \) here represents the index of the strip in each scale \( S \). Here, the visual features after being pooled are represented by \( f \), and the formula of global pooling is

\[
f_{s,t} = AP(V_{s,t}) + MP(V_{s,t}),
\]

where MP and AP respectively represent max-pooling and average-pooling. MP is good for extracting highly discrimi-
inative information, while AP is good for extracting global information. The integration of the two can not only extract the global information, but also extract the local information. Secondly, the first layer FC maps the feature \( f \) to the discrimination subspace, the constraint on this layer is triplet loss. The second layer FC is the discrimination space of the first layer FC output for ID classification.

Then, we need to consider the design and definition of constraint functions. Most of exist methods use ID as classifier label, in this case, the formation formula of cross entropy loss function is as follows:

\[
L_c = - \sum_{m=1}^{M} \sum_{i}^{S} CE(y'_i, y_i),
\]

where \( M \) is the batch size, \( S \) is the number of blocks finally divided vertically, \( y_i \) is the true value label, \( y'_i \) is the predicted value of the image output, \( CE \) is cross entropy loss. The classifier is implemented as a fully connected layer, and the output length is the total number of training IDs. ID classifier can be expressed as the cross entropy error between the prediction labels and the truth labels.

However, the use of the cross entropy loss function alone does not ensure that the learning features are discriminative. Especially for the gait silhouette information, its variation is greatly affected by walking conditions, that is, the same ID varies greatly under different walking conditions. Therefore, we introduce the triplet loss function, which is used to learn the end-to-end deep metric, which is used to pull the samples of the same ID closer in the feature space and push the samples of different IDs further in the feature space.

Given an image triplet, i.e., anchor \( a \), positive \( p \) and negative \( n \), the formation formula of triplet loss function is as follows:

\[
L_t = [m + d_{j,a,n}^p]_+ \tag{4}
\]

\[
d_{j,a,n}^p = D(f(x_j^p), f(x_a^p)) - D(f(x_j^p), f(x_n^p)) \tag{5}
\]

among them, \( f(x_j^p), f(x_a^p) \) and \( f(x_n^p) \) are features from anchor sample, positive sample and negative sample respectively. \( D(\cdot, \cdot) \) is the distance between features measured by Euclidean distance, \( m \) is the edge amplitude of the distance between positive and negative samples of the triplet function.

According to the above analysis, under the same gait recognition framework, the triplet loss and cross entropy loss are fused. The final objective function is formulated as follows:

\[
L = \lambda L_c + L_t, \tag{6}
\]

where \( L_c \) means the cross-entropy loss and \( L_t \) means the triplet loss. \( \lambda \in [0, 1] \) weights those two loss function. When \( \lambda \) is set to 0, only the \( L_t \) is considered. When \( \lambda \) is set to 1, \( L_c \) and \( L_t \) are considered. The effect of \( \lambda \) is discussed in Sec IV-D.

**IV. EXPERIMENT**

Our experiments consist of three parts. The first part is the experimental settings related to this method. The second part compares the proposed method with other advanced algorithms on CASIA-B and OU-LP-Bag datasets. The third part is ablation experiments on CASIA-B dataset.

**A. SETTINGS**

We use two datasets, CASIA-B [32] and OU-LP-Bag [33], to evaluate our method. CASIA-B and OU-LP-Bag are two public gait datasets. The details of these datasets are as follows.

CASIA-B [32] is a large-scale, multi-view gait dataset, samples as shown in Fig. 3. There are 124 subjects, and 11 views\((0°, 18°, 36°, \ldots, 180°)\) per subject, it contains normal walking conditions NM), wearing with coats(CL), walking with bags(BG). Namely, each subject has three walking conditions, including 6 video sequences per person under NM conditions, 2 video sequences per person under CL conditions and 2 video sequences per person under BG conditions. There are 11 views per walking condition, so each subject has 110 video sequences. The dataset is divided into three settings according to the data size of the training set, in first set, the training sets are the first 24 subjects, and the test sets are the rest 100 subjects, which are called small sample training (ST). In second set, the training sets are the first 62 subjects, and the test sets are the rest 62 subjects, which are called medium sample training (MT). In third set, the training sets are the first 74 subjects, and the test sets are the rest 50 subjects, which are called large sample training (LT). In the test set, the three settings are the same, the first 4 sequences under NM conditions are as gallery, the rest 6 sequences remain as probe.

OU-LP-Bag [33] is the world’s largest gait recognition database at present that focuses on carrying real-life carried objects. It consists of 62,528 subjects with seven distinct labels for the carrying status, samples as shown in Fig. 3. Each subject contains three sequences (A1, A2, and A3), including 60,450 subjects with the A1 sequence, 58,859 subjects with the A2 sequences and 58,709 subjects with the A3 sequences. The A1 sequence contains with or without carried objects, and the A2 and A3 sequences are both without carried objects. According to the literature settings [33], 58,199 subjects who had samples in the A1 sequence and in the A2 or A3 sequences for each subject were selected from 62,528 subjects in the dataset. Then, the 58,199 subjects was randomly divided into two sets by subject id: the training set consisting of 29,097 subjects and test set consisting of 29,102 subjects. The test sets was divided into gallery and probe set. In order to better reflect the real scene, there are cooperative settings and uncooperative settings. The cooperative setting is to leave the A2 or A3 sequences in the gallery and the A1 sequence in the probe. The uncooperative setting is to leave the A1 and either the A2 or A3 sequences in the gallery and the rest sequences(A2 or A3) kept in the probe.

**B. IMPLEMENTATION DETAILS**

The proposed method is implemented with PyTorch, and trained on a TITAN Xp GPU. During training, gait silhouettes are gray-scale image, and the input image size is 44 × 64.
The silhouettes are aligned based on methods in [2]. Adam is chosen as an optimizer [40]. After two down sampling operations, it becomes 16 × 11. A batch of P × K size is selected from the training set, where p represents the number of people and K represents the number of training samples each person has in the batch. Here P is set as 8 and K is set as 16. The m for triplet loss is set to 0.2. The learning rate is set to 1e−4. In CASIA-B, due to the different size of training set data, for ST, MT and LT, we conducted 50,000, 60,000 and 80,000 iterations training on our model respectively. In OU-LP-bag, we conducted 50,000 iterations training. During testing, given a query and gallery, both extract frame level features by visual representation module, and aggregate them as the global features of the sequence. Then it is sent to the high-level semantic metric module to get the final feature representation. Finally, each query is compared with all samples of gallery use Euclidean distance to calculate Rank-1 recognition accuracy.

C. EXPERIMENTAL RESULTS

1) RESULT ON CASIA-B

Table 2 reports the performance of our approach and existing state-of-the-art methods on CASIA-B, respectively. The results are averaged over the 11 views and excluded the same perspective. For example, the accuracy of the 18° probe is the average of 10 gallery views except 18° gallery.

Gait information includes not only the gait information parallel to the walking direction, but also the gait information perpendicular to the walking direction.
perpendicular to the walking direction. Although it is easy to lose part of the gait information in the parallel view angle (90°) and the vertical view angle (0° and 180°). From Table 2, the proposed method still has good performance. Compared with the current optimal Gaitset [6] algorithm, the viewing angles of 90°, 0°, and 180° increased by an average of 1.8%, 2.4% and 2.7%, respectively. At the same time, it improved by 0.8% across 11 views.

The proposed method has been improved on the small sample training (ST) with only 24 objects under three walking conditions (NM, BG, CL). Outperforming most of the recent work, e.g., AE [36], CMCC [34], CNN-LB [35] and GaitSet [6]. It has been improved by 2.8%, 4.4% and 0.7% compared with the current optimal of GaitSet [6] method. The reason is that because our model takes the input as a set, our model obtain the number of silhouettes more than base on gait templates (AE [36], CMCC [34] and CNN-LB [35]) in a same batch. Therefore, our model get more temporal and spatial information. Then, compared with the GaitSet [6] more discriminative gait features are acquired through metric learning, and a good training model is obtained. In addition, the overall average Rank-1 accuracy of 11 viewing angles increased by 2.6% under three walking conditions. In addition, in order to obtain more fair and convincing experimental results, We prepared 6 groups of experiments and averaged them, the results are shown in Table 3.

Appearance change is a big challenge in gait recognition. Our method has good performance in the BG sequence on CASIA-B. Table 4 shows that our method has achieved very good results in BG subsequence. It is better than the current optimal algorithm in ST, MT and LT, and it is the best in 11 views and average Rank-1 accuracy is 4.4%, 1.3% and 1.5% higher than other models, respectively. This also shows that gait features have strong invariance, which is for all major gait variations.

Some literatures [7], [35], [37], [38] also pay attention to the cross-view conditions. However, when the size of training set increases and the data interference increases, our model is also the best in accuracy. Table 5 shows that when the larger sample training (LT) with 74 objects, the walking conditions are NM, BG and CL, and cross angle of view (0°, 54°, 90°, 126°), the average accuracy of our model achieves the gain of 7%.

At the same time, some literatures [7], [8], [35], [41], [42] also pay attention to the appearance changes. According to the relevant literature settings, the training set is 54° to 144° in the case of BG, and different test sets contain different view, combinations of probes and appearance conditions (BG). Table 6 shows that our performance on the BG subset is out performance the current state-of-the-art JUCNet [8] method. It is worth noting that owing to the challenge of appearance change, that is a significant performance by a big margin between BG and NM in all methods except our method, which also demonstrate that our gait features have strong invariance to all major gait changes.

**TABLE 3.** Averaged Rank-1 accuracies on CASIA-B (ST) under three different experimental settings, that is 6 groups.

| Sets     | Methods | Mean |
|----------|---------|------|
| NM#5-6   | ViDP[1] | -    |
|          | CMCC[34]| -    |
|          | CNN-LB[35]| -   |
|          | GaitSet[6]| 79.5 |
|          | Ours    | 91.9 |
| BG#1-2   | GaitSet[6]| 63.6 |
|          | Ours    | 71.4 |
| CL#1-2   | GaitSet[6]| 40.9 |
|          | Ours    | 41.5 |

**TABLE 4.** Rank-1 accuracy of BG subsequence on CASIA-B.

| Gallery NM#1-4 | Probe | 0° | 18° | 36° | 54° | 72° | 90° | 108° | 126° | 144° | 162° | 180° | Mean |
|----------------|-------|-----|-----|-----|-----|-----|-----|------|------|------|------|------|------|
| ST BG#1-2      | GaitSet[6]| 55.8 | 70.5 | 76.9 | 75.5 | 60.7 | 63.4 | 68.0 | 75.8 | 76.2 | 70.7 | 52.5 | 68.6 |
|                | Ours    | 62.6 | 73.7 | 78.3 | 77.4 | 73.4 | 76.6 | 72.4 | 79.4 | 78.5 | 73.0 | 58.1 | 73.0 |
| MT BG#1-2      | ABE[36] | 29.8 | 37.7 | 39.2 | 40.5 | 43.8 | 37.5 | 43.0 | 42.7 | 36.3 | 30.6 | 28.5 | 37.2 |
|                | MGAN[12]| 48.5 | 58.9 | 59.7 | 58.0 | 53.7 | 49.8 | 54.0 | 61.3 | 59.5 | 55.9 | 43.1 | 54.7 |
|                | GaitSet[6]| 79.9 | 89.8 | 91.2 | 86.7 | 81.6 | 76.7 | 81.0 | 88.2 | 90.3 | 88.5 | 73.0 | 84.3 |
|                | Ours    | 79.4 | 88.1 | 92.0 | 90.0 | 84.2 | 80.2 | 83.9 | 89.1 | 92.6 | 87.1 | 74.6 | 85.6 |
| LT BG#1-2      | CNN-LB[35]| 64.2 | 80.6 | 82.7 | 76.9 | 68.4 | 63.1 | 68.0 | 76.9 | 82.2 | 75.4 | 61.3 | 72.4 |
|                | GaitNet[7]| 83.0 | 87.5 | 88.3 | 93.3 | 82.6 | 78.9 | 89.5 | 91.0 | 86.1 | 81.2 | 85.6 | 85.7 |
|                | GaitSet[6]| 83.8 | 91.2 | 91.8 | 88.8 | 83.3 | 81.0 | 84.1 | 90.0 | 92.2 | 94.4 | 79.0 | 87.2 |
|                | Ours    | 86.6 | 92.7 | 93.1 | 90.4 | 85.3 | 81.6 | 86.8 | 92.4 | 93.8 | 92.1 | 81.4 | 88.7 |
TABLE 7. The Rank-1, Rank-5, for cooperative (Coop) and uncooperative (Uncoop) settings of different methods on the OU-LP-Bag dataset.

| Methods   | Rank-1(%) | Rank-5(%) |
|-----------|-----------|-----------|
| DM[43]    | 17.7      | 15.9      | 23.4      | 20.5      |
| GEINet[44] | 22.3      | 18.5      | 32.5      | 26.9      |
| RSV[45]   | 24.7      | 18.3      | 35.6      | 27.6      |
| GERP[46]  | 38.5      | 31.2      | 50.9      | 42.2      |
| PCA_LDA[47] | 40.8      | 31.4      | 53.0      | 41.3      |
| SIAME[16] | 49.8      | 50.3      | 69.7      | 70.7      |
| JITN[48]  | 74.4      | 74.0      | -         | -         |
| ICDNet[49]| 87.0      | 86.5      | -         | -         |
| Ours      | 93.2      | 92.6      | 94.9      | 93.3      |

2) RESULT ON OU-LP-BAG

Following [33] and report the Rank-1 accuracy on the OU-LP-Bag dataset, which is the world’s largest gait recognition database at present that focuses on carrying real-life carried objects. We evaluate our method on Rank-1 and Rank-5 are reported in Table 7. It shows that, under the cooperative and uncooperative setting, our method achieves the best Rank-1 accuracy of 93.2% and 92.6%, outperforming most of the recent work by large margins in Rank-1 accuracy, e.g., under the cooperative setting, PCA_LDA [47], SIAME [16], JITN [48] and ICDNet [49] by 52.4%, 43.4%, 18.8% and 6.2% in Rank-1 accuracy, respectively. Note that, JITN introduces a complex joint intensity transformer network and uses contrastive loss and triplet loss, but in fact, contrast loss only constrains the features of pairs within a class as close as possible, while pairs between classes have features as far as possible. Compared with JITN and ICDNet method, our method achieves excellent performance over recently works. In all the evaluation algorithms, our models all achieved excellent performance. This shows that the model has strong robustness in learning gait features.

D. ABLATION EXPERIMENTS

1) IMPACT OF METRIC LEARNING

To evaluate the effectiveness of the proposed method on CASIA-B, ablation experiments are conducted on three conditions with 11 views: ST. We consider four setting with different function: $L_c$, $L_t$, $L_c + L_t$ and $\lambda L_c + L_t$. All experimental results are shown in Table 8. The reason the experiment was set up is that whether the proposed method is suitable for feature learning of small samples, whether it performs better in all three walking conditions like NM, BG, CL, and whether the fusion of loss functions is effective. Also, when only triplet loss is added and only cross entropy loss function is added, triplet loss is more effective than cross entropy loss. From the results, we can obtain some key observations. When combined with training, the ratio in order to optimize the joint training, the parameter $\lambda$ is added to adjust the weight of cross entropy loss. By adjusting $\lambda$, the overall average Rank-1 accuracy is improved by 1.7% compared with that the parameter $\lambda$ as 0. In terms of NM, BG and CL settings, the average Rank-1 accuracy is increased by 2.1%, 2.5% and 0.5%, respectively. Specially, in BG sequence, the Rank-1 accuracy was improved by 2.5%. This indicates that the proposed method suitableness for feature learning of small samples, performs better in three walking conditions, and fuses loss functions is effective.

2) IMPACT OF $\lambda$

We evaluate the influence of $\lambda$ on Rank-1 accuracy under the ST, MT, and LT setting of CASIA-B dataset of BG condition. The impact of the parameter $\lambda$ is shown in Fig.4. Notice that, when $\lambda$ is set to 0, we only consider the triplet loss; when $\lambda$ is equal to 1, triplet and cross-entropy loss are considered. It can be observed that When $\lambda$ is set to 0.01, the overall average Rank-1 accuracy is improved by 1.4% compared with only consider the triplet loss, as the same time, in terms of ST, MT and LT settings, the average Rank-1 accuracy is increased by 1.4%, 1.3% and 1.5% respectively. It’s clear that $\lambda$ is set to 0.01 achieves the best performance. Therefore, we set $\lambda$ as 0.01 in the all experiments.

3) IMPACT OF POOLING

In this paper, in order to maintain the global relation with the identification and preserve the discriminative part, we use both average pooling (AP) and max pooling (MP). In addition, the proposed method uses AP and MP together, not separately. Table 9 shows the results with different pooling strategies. It can be observed that MP performs better than AP.

TABLE 8. Ablation experiments on CASIA-B with ST, MT and LT settings. The results were 11 views and the average Rank-1 accuracy.

| Gallery | NM#1-4 | 0°-180° | Mean |
|---------|--------|---------|------|
| Probe   | 0°     | 18°     | 36°  | 54°  | 72°  | 90° | 108°  | 126° | 144° | 162° | 180° | |
| L_c     | 38.8   | 67.2    | 81.0 | 81.4 | 72.1 | 66.4 | 73.1  | 77.8 | 73.3 | 62.0 | 35.4 | 66.2 |
| L_t     | 68.8   | 80.8    | 84.9 | 86.3 | 79.5 | 74.6 | 80.7  | 86.0 | 87.6 | 81.8 | 64.6 | 80.0 |
| L_c + L_t | 66.3 | 81.9    | 89.9 | 88.1 | 82.2 | 75.9 | 81.9  | 86.7 | 86.9 | 81.7 | 65.4 | 80.6 |
| $\lambda L_c + L_t$ | 69.0 | 84.8    | 90.5 | 88.5 | 82.9 | 78.5 | 82.5  | 88.4 | 89.8 | 83.0 | 67.6 | 82.3 |
| ST      |        |         |      |      |      |      |       |      |      |      |     |
| NM#5-6 | L_c    | 26.4    | 48.0 | 63.0 | 67.2 | 59.2 | 53.2  | 57.7 | 62.7 | 54.1 | 42.8 | 25.0 |
|         | L_t    | 57.0    | 69.8 | 74.5 | 73.3 | 68.7 | 62.6  | 66.4 | 75.7 | 75.6 | 69.4 | 55.3 |
|         | $L_c + L_t$ | 56.5 | 69.9  | 76.8 | 74.9 | 70.0 | 62.4  | 67.1 | 75.8 | 75.2 | 68.3 | 55.8 |
|         | $\lambda L_c + L_t$ | 62.6 | 73.7  | 78.3 | 77.4 | 73.4 | 72.6  | 72.4 | 79.4 | 78.5 | 73.0 | 58.1 |
| BG#1-2 | L_c    | 12.6    | 23.0 | 31.2 | 30.0 | 26.6 | 22.3  | 25.3 | 26.5 | 21.2 | 17.5 | 11.0 |
|         | L_t    | 30.4    | 44.6 | 49.8 | 46.9 | 43.2 | 39.5  | 41.5 | 43.3 | 41.7 | 32.8 | 26.5 |
|         | $L_c + L_t$ | 32.4 | 43.4  | 47.5 | 46.9 | 43.4 | 38.3  | 42.3 | 44.2 | 40.4 | 32.0 | 27.1 |
|         | $\lambda L_c + L_t$ | 32.9 | 46.6  | 49.6 | 46.8 | 44.9 | 40.6  | 44.8 | 42.1 | 35.1 | 29.1 | 41.6 |
in the most cases. The reason is that AP takes into account all the locations of a particular part, and all locations contribute equally to the final part representation. Therefore, unrelated background patterns can easily affect the recognition ability of the representations generated by AP. Conversely, the global MP only retains the maximum response value for the local view. We think these two pooling strategies are complementary in generating feature representations from global and local competition. Therefore, we integrate them into a unified model to take full advantage of these two strategies. Experimental results in Table 9 demonstrate that mixing the two pooling strategies achieves better results compared with using either of them.

4) IMPACT OF GLOBAL FEATURE LAYERS
To evaluate the impact of global feature layers, G1 and G2, we conducted experiment on CASIA-B dataset. The results are shown in Table 10. It can be observed that the combination of G1 and G2 can achieve better performance. This is because shallow features pay more attention to fine-grained, while deep features contain more global coarse-grained information. Thus, we propose a multi-level CNN visual representation extraction module, It has a similar structure to the CNN in the main pipeline and incorporates the layer features extracted in different layers into G1-G2.

5) IMPACT OF FULLY CONNECTION LAYERS
In order to help the deep network focus on features with different sizes to gather both local and global information. We improve the high-level semantic metric module to make it adapt better for gait recognition task. We design two fully connection (FC) layers, the first FC layer for each pooled feature to map it into the discriminative space, the second FC layer is for ID classification. according to [31], 1 × 1 convolutional layers are used to reduce the dimensions of the column feature maps. However, in this paper, it need a classifier after the pooling. Obviously, 1 × 1 convolutional layer cannot do this job, so the full connection layer is used here. In order to better understand the effect of the two FCs, the following ablation experiments were conducted, as shown in Table 11. It can be observed that two FCs performs better than one FC in a large margin.

V. CONCLUSION
This paper presents a gait recognition method based on joint metric learning and hierarchical network. Compared with the existing methods, our method focus on feature space optimization which can extract spatial and temporal information more effectively, and has a great improvement in a discriminative gait feature. The current perspective and the change of carrying objects are the two challenges in gait recognition, as the main factor affecting the rate of gait recognition. In this paper they have been effectively improved here. The experiments of two benchmark gait datasets (open standard datasets) show that compared with other most advanced algorithms, it achieves the highest recognition accuracy and shows well flexibility in complex environments. At present, gait recognition still has problems such as gait occlusion and dress change that need to be solved. Therefore, in the future, we will focus on these problems that need to be solved and extend the applications of the gait recognition, such as person Re-Identification and behavior recognition.

REFERENCES
[1] M. S. Nixon, T. Tan, and R. Chellappa, “Human identification based on gait,” The Kluwer International Series on Biometrics, vol. 4, New York, NY, USA: Springer, 2006.
[2] N. Takemura, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “Multi-view large population gait dataset and its performance evaluation for cross-view gait recognition,” IPSJ Trans. Comput. Vis. Appl., vol. 10, p. 4, Dec. 2018.
A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and S. Chopra, R. Hadsell, and Y. LeCun, “Learning a similarity metric

D. Das, A. Agarwal, P. Chattopadhyay, and L. Wang, “RGait-NET: An

Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning

H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, “Deep relative distance

K. Sohn, “Improved deep metric learning with multi-class N-pair loss objective,” in Proc. Adv. Neural Inf. Process. Syst., Am. Conf. Neural Inf. Process. Syst., Barcelona, Spain, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., Dec. 2016, pp. 1849–1857.

H. Liu, Y. Tian, Y. Wang, L. Pang, and T. Huang, “Deep relative distance learning: Tell the difference between similar vehicles,” in Proc. CVPR. Las Vegas, NV: IEEE Computer Society, 2016, pp. 4004–4012.

B. G. V. Kumar, G. Carneiro, and I. D. Reid, “Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions,” CoRR, vol.abs/1512.09272, pp. 1–10, Dec. 2015.

Y. Wen, K. Zhang, Z. Li, and Y. Qiao, “A discriminative feature learning approach for deep face recognition,” in Proc. ECCV, in Lecture Notes in Computer Science, vol. 9911, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Amsterdam, The Netherlands: Springer, 2016, pp. 499–515.

V. Ustinova and V. S. Lampitsky, “Learning deep embeddings with histogram loss,” in Proc. Adv. Neural Inf. Process. Syst., Am. Conf. Neural Inf. Process. Syst., Barcelona, Spain, D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, Eds., Dec. 2016, pp. 4170–4178.

A. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in Proc. CVPR. Columbus, OH, USA: IEEE Computer Society, 2014, pp. 512–519.

T. Wang, S. Gong, X. Zhu, and S. Wang, “Person re-identification by discriminative selection in video ranking,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 12, pp. 2501–2514, Dec. 2016.

W. Li, X. Zhu, and S. Gong, “Harmonious attention network for person re-identification,” in Proc. CVPR. Salt Lake City, UT: USA: IEEE Computer Society, 2018, pp. 2285–2294.

A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Commun. ACM, vol. 60, no. 6, pp. 84–90, Jun. 2017.

K. Simonany and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. 3rd Int. Conf. Learn. Represent. (ICLR), San Diego, CA, USA, Y. Bengio and Y. LeCun, Eds., May 2015. [Online] Available: http://arxiv.org/abs/1409.1556.

C. Szegedy, W. Liu, J. Y. P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR). Boston, MA, USA: IEEE Computer Society, Jun. 2015, pp. 1–9.

Y. Fu, Y. Wei, Y. Zhou, H. Shi, G. Huang, X. Wang, Z. Yao, and T. S. Huang, “Horizontal pyramid matching for person re-identification,” in Proc. AAAI. Honolulu, HI: USA: AAAI Press, 2019, pp. 8295–8302.

S. Yu, D. Tan, and T. Tan, “A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition,” in Proc. 18th Int. Conf. Pattern Recognit. (ICPR). Hong Kong: IEEE Computer Society, 2006, pp. 441–444.

M. Z. Uddin, T. T. Ngo, Y. Makihara, N. Takemura, X. Li, D. Muramatsu, and Y. Yagi, “The OU-ISIR large population gait database with real-life carried object and its performance evaluation,” IPSJ Trans. Comput. Appl., vol. 10, no. 1, p. 5, Dec. 2018.

W. Kasukunrnn, Q. Wu, J. Zhang, H. Li, and L. Wang, “Recognizing gait across views through correlated motion clustering,” IEEE Trans. Image Process., vol. 23, no. 2, pp. 696–709, Feb. 2014.

Z. Wu, Y. Huang, L. Wang, X. Wang, and T. Tan, “A comprehensive study on cross-view gait based human identification with deep CNNs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 2, pp. 209–226, Feb. 2017.

S. Yu, H. Chen, Q. Wang, L. Shen, and Y. Huang, “Invariant feature extraction for gait recognition using only one uniform model,” Neurocomputing, vol. 239, pp. 81–93, May 2017.

W. Kasukunniran, Q. Wu, J. Zhang, and H. Li, “Support vector regression for multi-view gait recognition based on local motion feature selection,” in Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., IEEE Computer Society, Jun. 2010, pp. 974–981.

W. Kasukunrnn, Q. Wu, H. Li, and J. Zhang, “Multiple views gait recognition using multi-view transformation model based on optimized energy image,” in Proc. IEEE 12th Int. Conf. Comput. Vis. Workshops, ICCV Workshops, IEEE Computer Society, Sep. 2009, pp. 1058–1064.

K. Bashar, T. Xiang, and S. Gong, “Cross view gait recognition using correlation strength,” in Proc. Brit. Mach. Vis. Conf., F. Labrosse, R. Zwiggelaar, Y. Liu, and B. Tiddeman, Eds. Durham, U.K.: British Machine Vision Association, 2010, pp. 1–11.

D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in Proc. ICLR. Y. Bengio and Y. LeCun, Eds., 2015, pp. 1–15.

X. Chen, J. Weng, W. Lu, and J. Xu, “Multi-gait recognition based on attribute discovery,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 7, pp. 1697–1710, Jul. 2018.

H. Hu, “Enhanced Gabor feature based classification using a regularized locally tensor discriminant model for multiview gait recognition,” IEEE Trans. Circuits Syst. Video Technol., vol. 23, no. 7, pp. 1274–1286, Jul. 2013.

H. Iwama, M. Okumura, Y. Makihara, and Y. Yagi, “The OU-ISIR large population gait database comprising the large population dataset and performance evaluation of gait recognition,” IEEE Trans. Inf. Forensics Security, vol. 7, no. 5, pp. 1511–1521, Oct. 2012.

K. Shiraga, Y. Makihara, D. Muramatsu, T. Echigo, and Y. Yagi, “GEINet: View-invariant gait recognition using a convolutional neural network,” in Proc. Int. Conf. Biometrics (ICBM), Halmstad, Sweden, Jun. 2016, pp. 1–8.

O. Chapelle and S. S. Keerthi, “Efficient algorithms for ranking with SVMs,” Inf. Retr., vol. 13, no. 3, pp. 201–215, Jun. 2010.

X. Li, Y. Makihara, C. Xu, D. Muramatsu, Y. Yagi, and M. Ren, “Gait energy response function for clothing-invariant gait recognition,” in Proc. ACCV, in Lecture Notes in Computer Science, vol. 10112, S. Lai, V. Lepetit, K. Nishino, and Y. Sato, Eds., 2017, pp. 257–272.
[47] N. Otsu, “Optimal linear and nonlinear solutions for least-square discriminant feature extraction,” in Proc. 6th Int. Conf. Pattern Recognit., Munich, Germany, 1982, pp. 557–560.

[48] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, “Joint intensity transformer network for gait recognition robust against clothing and carrying status,” IEEE Trans. Inf. Forensics Security, vol. 14, no. 12, pp. 3102–3115, Dec. 2019.

[49] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren, “Gait recognition via semi-supervised disentangled representation learning to identity and covariate features,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 13309–13319.

[50] C. Xu, Y. Makihara, X. Li, Y. Yagi, and J. Lu, “Cross-view gait recognition using pairwise spatial transformer networks,” IEEE Trans. Circuits Syst. Video Technol., early access, Feb. 21, 2020, doi: 10.1109/TCSVT.2020.2975671.

HUANHUAN XU is currently pursuing the M.S. degree with the School of Computer Science and Technology, Tiangong University. Her research interests include gait recognition and person re-identification.

YUQIAN LI is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering, Tianjin University, China. His current research interests include computer vision and machine learning.

**XUEMEI SUN** received the Ph.D. degree from Tianjin University, Tianjin, China, in 2005. She has been a Professor with the School of Computer Science and Technology, Tiangong University. She has published over 20 papers in refereed conferences and journals. One of the representative papers, Optimized Deployment of Wireless Sensor Networks Based on Ant Colony Algorithm Culture, was published, in 2015, in the Journal of Applied Mathematics and Computing. Her current research interests include cloud computing, artificial intelligence, computer vision, and big data.

**SHENGJIN WANG** (Senior Member, IEEE) received the B.E. degree from Tsinghua University, Beijing, China, in 1985, and the Ph.D. degree from the Tokyo Institute of Technology, Tokyo, Japan, in 1997. From May 1997 to August 2003, he was a member of the Senior Research Staff with the Internet System Research Laboratories, NEC Corporation, Nara, Japan. Since September 2003, he has been a Professor with the Department of Electronic Engineering, Tsinghua University, where he is currently the Director of the Research Center for Media Big-data Cognitive Computing. He has published more than 100 articles. He holds more than 20 patents. His current research interests include computer vision, pattern recognition, object detection, advanced driving assistant systems, and deep learning.