Deep Ranking Model by Large Adaptive Margin Learning for Person Re-identification

Jiayun Wang[a], Sanping Zhou[a], Jinjun Wang[a,*], Qiqi Hou[a]

[a]*The institute of artificial intelligence and robotic, Xi'an Jiaotong University, Xianning West Road No.28, Shaanxi, 710049, P.R. China*

Abstract

Person re-identification aims to match images of the same person across disjoint camera views, which is a challenging problem in video surveillance. The major challenge of this task lies in how to preserve the similarity of the same person against large variations caused by complex backgrounds, mutual occlusions and different illuminations, while discriminating the different individuals. In this paper, we present a novel deep ranking model with feature learning and fusion by learning a large adaptive margin between the intra-class distance and inter-class distance to solve the person re-identification problem. Specifically, we organize the training images into a batch of pairwise samples. Treating these pairwise samples as inputs, we build a novel part-based deep convolutional neural network (CNN) to learn the layered feature representations by preserving a large adaptive margin. As a result, the final learned model can effectively find out the matched target to the anchor image among a number of candidates in the gallery image set by learning discriminative and stable feature representations. Overcoming the weaknesses of conventional fixed-margin loss functions, our adaptive margin loss function is more appropriate for the dynamic feature space. Using the PRID2011, Market1501, CUHK01 and 3DPeS, we extensively conduct comparative evaluations to demonstrate the advantages of the proposed method over the state-of-the-art approaches in person re-identification.

Keywords: Person Re-identification, Deep Ranking Model, Metric Learning

Figure 1: The challenges to person re-identification in public space, in which the first row shows the training samples in camera A and the second row represents the training samples in camera B. Specially, (a) shows variations in lighting condition, (b) plots low resolution images, (c) illustrates variations of body pose, (d) denotes occlusions among pedestrians and (e) represents variations in view angle.

1. Introduction

Person re-identification (re-id) is a fundamental task in video surveillance and has attracted much attention in the visual recognition community [1–4]. Given one single shot or multiple shots of a target, it aims to match the same person among a set of gallery candidates captured from a disjoint camera view [5]. Since the matching results are ranked by the possibility of being the same identity, it is called "ranking" model. Despite years of efforts from researchers, the problem is still extremely challenging due to large appearance variations caused by light conditions, view angles, mutual occlusions and body poses, as shown in Fig. 1. In addition, pedestrian images captured by the surveillance system are usually in low resolution, which makes many visual details, such as the facial components, indistinguishable. Therefore, different individuals may look similar in appearance.

To address these challenges, extensive work has been reported in the past few years, and can be roughly divided into two classes [6, 7]: *feature representation* and *distance metric*. The feature representation based methods aim at constructing a discriminative visual descriptor that is robust to distinguish the different persons in disjoint camera views. Representative visual descriptors include the ensemble of local feature (ELF) [8], local maximal occurrence (LOMO) [9], and local binary pattern (LBP) [10]. The distance metric based methods aim at seeking

*Corresponding author: Tel.: +86-29-83395146; Fax: +86-29-83395175;
Email address: jinjun@mail.xjtu.edu.cn. (Jinjun Wang)

a proper similarity measurement by metric learning based on a group of labeled training data. Typical similarity measurements include the adaptive decision function (LADF) [11], large margin nearest neighbor (LMNN) [12] and pairwise constraint component analysis (PCCA) [13].

Recently, deep learning based methods are becoming increasingly popular in the person re-identification application [14–18] because they can incorporate *feature representation* and *distance metric* into an integrated framework. The feature representation and distance metric are performed as two different components: (1) a deep CNN to extract feature representations from the input images, and (2) a designed metric to back-propagate gradients of the loss function. Benefiting from the powerful representation capability of the deep CNN, the deep learning based methods have achieved state-of-the-art performance on the benchmark datasets of person re-identification.

A number of deep learning based methods apply the conventional contrastive loss or the triplet loss function to learn a fixed margin between the intra-class samples and inter-class samples, which do not suit the dynamic feature space well. These two loss functions also have some other drawbacks, as discussed in Section 3.3. Besides, as a result of lacking labeled training data, a person re-identification dataset is usually not huge in size, which highly limits the generalization ability of models on the testing data. To alleviate these problems, we consider two notable aspects in our method: (1) the deep constraint should be adaptive to the dynamic feature space, and (2) the deep architecture should be more shallow and expressive, in order to cope with small or medium sized dataset. In this paper, we propose a novel deep ranking model with feature learning and fusion by learning a large adaptive margin between the intra-class samples and inter-class samples for person re-identification. First, in order to generate discriminative feature representations for different individuals, we build a part-based deep CNN to learn feature representations from multiple perspectives of a pedestrian, in which different body parts are discriminatively learned in the convolutional layers and then fused in the fully connected layers. Secondly, we propose a novel distance metric to supervise the training of the deep CNN, in which the intra-class distance is minimized and the inter-class is maximized by adopting two large adaptive margin strategies, respectively.

In the following, we overview the main components of our method and summarize the advantages:

- By organizing the training images into a batch of pairwise samples, we propose a novel distance metric, in which the intra-class distance is minimized and the inter-class distance is maximized by preserving a large adaptive margin between them. Compared with the existing approaches, our method can

not only reduce the time of training the deep CNN, but also boost the ranking performance by adopting the adaptive margin strategy into the proposed distance metric.

- A novel part-based deep architecture is built to extract the discriminative feature representation of different body parts, which consists of three sub-networks, namely the global sub-network, local sub-network and fusion sub-network. Different body parts are first discriminately learned in the global sub-network and local sub-network, and then fused in the fusion sub-network. As a result, the finally learned feature representation is robust to the variations in light conditions, view angles, mutual occlusions and body poses.

- We conduct extensive experiments to evaluate various aspects of our method on the benchmark datasets, namely the PRID2011, Market1501, CUHK01 and 3DPeS, and the results outperform the state-of-the-art on all four datasets.

The rest of the paper is organized as follows. In Section 2, we briefly review the related works. Section 3 introduces the proposed large adaptive margin distance metric and the part-based deep neural network, followed by the optimization in Section 4. Section 5 demonstrates the organization of training samples. The experimental results and result analysis are presented in Section 6. Section 7 concludes the work.

2. Related Work

There has been extensive work to tackle the person re-identification problem, and most focus on feature representation, distance metric and deep learning. We present below some of the related works in terms of these three aspects.

Feature representation The feature representation based methods mainly focus on developing a discriminative visual descriptor, which is robust to the view angles, body poses, light conditions and mutual occlusions. For example, Zhao et al. [9] learned a mid-level filter from path cluster to achieve cross view invariance. Liao et al. [19] constructed a feature descriptor, which analyzed this horizontal occurrence of local features and maximized the occurrence to make a stable representation against viewpoint changes. Ma et al. [20] presented the person image via covariance descriptor which was robust to illumination change and background variations. Farenzena [21] et al. augmented maximally stable color regions with histograms for person representation. Zhao et al. [22] learned the distinct salience feature to distinguish the matched person from others. Chen et al. [23] employed a pre-learned pictorial structure model to localize the body parts more accurately.

Wu et al. [24] introduced a viewpoint-invariant descriptor, which took the viewpoint of the human into account by using what they called a pose prior, learned from training data. Kviatkovsky et al.[25] investigated the intra-distribution structure of color descriptor, which was invariant under certain illumination changes. Li et al. [26] matched person images observed in different camera views with complex cross-view transforms and applied it to person re-identification.

Distance metric The distance metric based methods aim to seek a proper similarity measurement, in which feature representations from the same person are closer than those from different ones. For example, Zheng et al. [27] proposed a relative distance learning method from a probabilistic prospective. Mignon et al.[28] learned a distance metric from sparse pairwise similarity constraints. Pedagadi et al. [29] utilized local Fisher Discriminant Analysis (LFDA) to map high dimensional features into a more discriminative low dimensional space. Xiong et al. [10] further extended the LFDA and several other metric learning methods by using kernel tricks and different regularizers. Nguyen et al. [30] measured the similarity of face pairs through cosine similarity, which was closely related to the inner product similarity. Loy et al. [31] modeled the person re-identification problem as an image retrieval task by considering the listwise similarity. Chen et al. [32] proposed a kernel based metric learning method to explore the nonlinearity relationship of samples in the feature space. Hirzer et al. [33] learned a discriminative metric by using relaxed pairwise constraints.

Deep learning The deep learning based methods aim to incorporate feature extraction and metric learning into an integrated framework, in which feature representation can be learned under the supervision of one similarity metric. [34–39] For example, Li et al. [40] proposed a novel filter pairing neural network (FPNN) to model body part displacements by using the patch matching layers to match the filter responses of local patches across views. Ahmed et al. [14] proposed an improved deep learning framework which takes pairwise images as inputs, and outputs a similarity value indicating whether the two input images depict the same person or not. Xiao et al. [41] proposed a domain guided dropout algorithm to improve the performance of deep CNN to extract robust feature representation for person re-identification. Yi et al. [42] constructed a siamese neural network to learn pairwise similarity, and used body parts to train the model. Ding et al. [15] applied the triplet loss to train the triplet deep framework for person re-identification. Wang et al.[43] proposed a unified triplet and siamese deep architecture which can jointly extract single-image and cross-image feature representations. Zhao et al. [9] used the local patch matching method that learns the mid-level filters to get the local discriminative features for person re-identification.

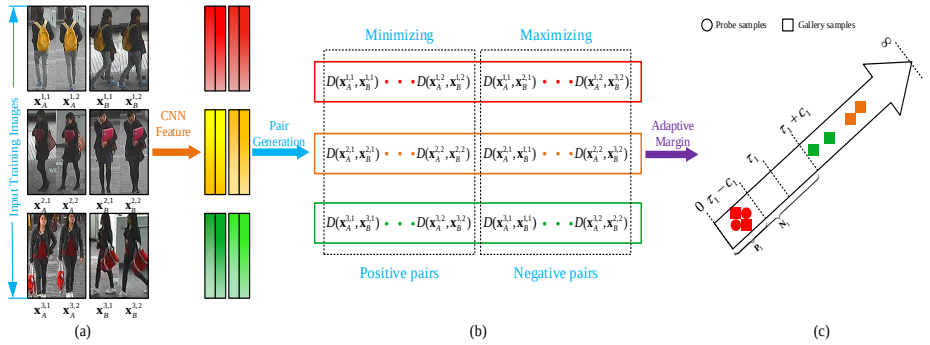

Figure 2: The illustration of deep ranking and large adaptive margin metric. (a) shows the training samples feeding into the deep CNN; (b) minimizes the intra-class distance and maximizes the inter-class distance in the feature space; (c) shows the large adaptive margin strategy used in our method while minimizing the intra-class distance and maximizing the inter-class distance.

3. Our method

In this section, we firstly introduce the *large adaptive margin* loss function, and then give a detailed illustration of our *deep neural network*. Fig. 2 shows the basic pipeline of our method: the deep neural network is designed to extract feature representations of the input images, and the large adaptive margin metric is used to compute the loss and back-propagate the gradients. As a result, the optimized deep CNN can learn robust feature representations, in which the intra-class distance is smaller than the inter-class distance.

3.1. Large adaptive margin

Let $\mathbf{X} = \{\mathbf{X}_n\}_{n=1}^{N}$ be the set of training samples, where $\mathbf{X}_n = \{\mathbf{X}_A^n, \mathbf{X}_B^n\}$ denotes the n^{th} person (class) of training samples. $\mathbf{X}_n$ consists of data from both camera A and camera B. In each camera view, the training samples can be represented by $\mathbf{X}_A^n = \{\mathbf{x}_A^{n,i}\}_{i=1}^{A_n}, \mathbf{X}_B^n = \{\mathbf{x}_B^{n,j}\}_{j=1}^{B_n}$, where $\mathbf{x}_A^{n,i}, \mathbf{x}_B^{n,j}$ are the i^{th} and j^{th} raw input images of the n^{th} class in both camera views, and A_n, B_n denote the corresponding image number. Note that $\mathbf{X}$ denotes a set of images, while $\mathbf{x}$ only refers one specific image. The goal of our deep model is to learn filter weights and biases that minimizes the ranking error from the output layer. A recursive function for an M-layer deep model can be formulated as follows:

$$\begin{aligned} \mathbf{X}_i^{(m)} &= \Psi_m(\mathbf{W}^{(m)} * \mathbf{X}_i^{(m-1)} + \mathbf{b}^{(m)}), \\ i = 1, 2, \cdots, N; m &= 1, 2, \cdots, M; \mathbf{X}_i^{(0)} = \mathbf{X}_i. \end{aligned} \tag{1}$$

where $\mathbf{W}^{(m)}$ denotes the filter weights of the m^{th} layer, $\mathbf{b}^{(m)}$ refers to the corresponding biases, $*$ denotes the convolution operation, $\Psi_m(\cdot)$ is an element-wise non-linear activation function such as rectified linear unit (ReLU), and $\mathbf{X}_i^{(m)}$ represents the feature maps generated at layer m for sample $\mathbf{X}_i$. For similarity, we simplify the parameters of the neural network as a whole and define $\mathbf{W} = \{\mathbf{W}^{(1)}, \cdots, \mathbf{W}^{(M)}\}$ and $\mathbf{b} = \{\mathbf{b}^{(1)}, \cdots, \mathbf{b}^{(M)}\}$.

For each pair of raw input images $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{j,s}$, we represent them as $f(\mathbf{x}_A^{i,k})$ and $f(\mathbf{x}_B^{j,s})$ at the output layer for the subsequent similarity comparison. The pairwise distance metric can be measured by computing the squared Euclidean distance between the final feature representations, which is defined as follows:

$$\mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}) = \|f(\mathbf{x}_A^{i,k}) - f(\mathbf{x}_B^{j,s})\|_2^2 \tag{2}$$

The smaller $\mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s})$ is, the more similar the two person images $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{j,s}$ are. Therefore, the definition formulates the person re-identification problem as a nearest neighbor search problem in the Euclidean space, which can be efficiently solved via similarity comparison algorithms.

3.2. Adaptive margin loss function

The proposed adaptive margin loss function is consisted of two terms: the similarity comparison term and the regularization term, which are formulated as follows:

$$\mathrm{L} = \mathrm{L_S}(\mathbf{X}, \mathbf{W}, \mathbf{b}) + \lambda R(\mathbf{W}, \mathbf{b}) \tag{3}$$

where $\mathrm{L_S}$ is the similarity comparison term, R represents the regularizer term, and λ is the lagrangian multiplier. Given a set of pairwise training samples, the similarity comparison term supervises the training process by minimizing the intra-class distance or maximizing the inter-class distance, while preserving a large adaptive margin between them. The regularizer term is used to smooth the parameters of the deep model, so as to avoid overfitting.

The similarity comparison term. In order to exploit the stable and discriminative feature representations at the output layer of deep CNN, we expect that there is a large margin between the positive pairs and negative pairs. In a mini-batch, $\mathbf{x}_A^{i,k}$ denotes the k^{th} capture of i^{th} person in camera view A, $\mathbf{X}_A^g$ denotes a set of person captures in camera view A that has property g. For the same person $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{i,s}$, the final distance $\mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{i,s})$ between $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{j,s}$ should be smaller than an adaptive up-margin $\mathcal{M}_p$. For different persons $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{j,s}$, the final distance $\mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s})$ between $\mathbf{x}_A^{i,k}$ and $\mathbf{x}_B^{j,s}$ should be larger than an adaptive down-margin

$\mathcal{M}_n$. The formulation can be represented as follows:

$$\begin{aligned} \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}) &< \mathcal{M}_p(\mathbf{x}_A^{i,k}, \mathbf{X}_B^{g}),\ i = j,\ i \neq g; \\ \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}) &> \mathcal{M}_n(\mathbf{x}_A^{i,k}, \mathbf{X}_B^{g}),\ i \neq j,\ i = g. \end{aligned} \tag{4}$$

Obviously, $\mathcal{M}_p$ is used to penalize the inter-class distances, and $\mathcal{M}_n$ is used to constrain the intra-class distances in the training process. $\mathcal{M}_p(\mathbf{x}_A^{i,k}, \mathbf{X}_B^{g})(\ i \neq g)$ indicates that the up-margin is calculated based on the anchor and its negative pairs. For this purpose, we formulate the adaptive margins as follows:

$$\begin{aligned} \mathcal{M}_p &= \frac{1}{\mu}(1 - \exp(-\mu d)); \\ \mathcal{M}_n &= \frac{1}{\gamma}\log(1 + \exp(\gamma s)). \end{aligned} \tag{5}$$

where the mean positive distance s and mean negative distance d are defined as follows:

$$\begin{aligned} s &= \frac{1}{N}\sum\nolimits_{s=1}^{N} \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}),\ \text{if } i = j. \\ d &= \frac{1}{N}\sum\nolimits_{d=1}^{N} \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}),\ \text{if } i \neq j. \end{aligned} \tag{6}$$

We consider the adaptive margin $\mathcal{M}_p$ and $\mathcal{M}_n$ as a nonlinear mapping of the average distances d and s. There are 2 advantages of the nonlinear mapping: 1) It is more suitable for the dynamic feature space. At the first several iterations of the training, the positive distance and the negative distance are similar. As depicted in Fig. 3, the nonlinear mapping penalizes the positive distance to get smaller, and the negative distance to get bigger, so that distinction between different identities can be achieved; when it comes to later iterations, as the difference between positive distance and negative distance gets more and more distinct, the nonlinear mapping effect diminishes significantly. Finally, $\mathcal{M}_p$ and $\mathcal{M}_n$ are close to the average positive distances and average negative distances, which can well avoid the over-fitting problem caused by the fixed margin strategy in the training process in earlier works. 2) The degree of the nonlinear mapping can be tuned by the hyper parameters. As shown in Fig. 3, the larger μ gets, the more powerful the nonlinear mapping becomes. Through tuning the 2 hyper parameters, our adaptive margin can adapt to different datasets. In short, the nonlinear mapping makes $\mathcal{M}_p$ get smaller than a moderate up-margin with the negative distances going up, and $\mathcal{M}_n$ get larger than a moderate down-margin with the positive distances going down.

Presetting $\mathcal{M}_p = \mathcal{M}_\tau - \mathcal{M}_c$ and $\mathcal{M}_n = \mathcal{M}_\tau + \mathcal{M}_c$, we can simplify the representation of constraint in Eq.(4), and the adaptive margin between the intra-class samples and inter-class sample can be enforced by using the following constraint:

$$\mathcal{M}_c - y_{ij}(\mathcal{M}_\tau - \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s})) < 0 \tag{7}$$

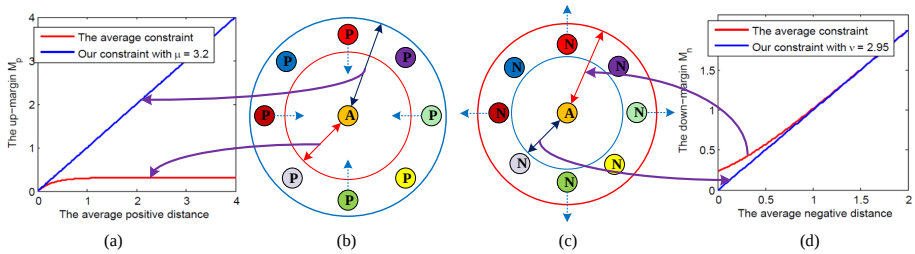

Figure 3: The illustration of two different nonlinear mapping strategies for the adaptive margin, in which (a) and (b) show the margin constrained to $\mathcal{M}_p$, (c) and (d) represent the margin constrained to $\mathcal{M}_n$. Circles refers to the adaptive margins, the color of the circle is correspondent to different nonlinear mapping degree (controlled by hyper parameters μ, γ) in (a) and (d). From the results, we can see that $\mathcal{M}_p$ will be penalized when the average negative distance is large and $\mathcal{M}_n$ will be compensated when the average positive distance is small. Besides, the degree of nonlinearization can be controlled by hyper parameters.

where $\mathcal{M}_\tau > \mathcal{M}_c$ and y_{ij} represents the relationship between the i^{th} and j^{th} identities, which is defined as follows:

$$y_{ij} = \begin{cases} 1, & \text{if } i = j; \\ -1, & \text{else.} \end{cases} \tag{8}$$

By applying the adaptive margin constraint in Eq. (6) to each positive pair and negative pair in the training set, the hinge-like loss function of $\mathrm{L_S}$ can be formulated as follows:

$$\mathrm{L_S} = \sum_{i,j=1}^{N} \max\{\mathcal{M}_c - y_{ij}(\mathcal{M}_\tau - \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s})), 0\} \tag{9}$$

The regularization term In order to smooth the parameters of the whole neural network, we define the following regularizer term, which can be formulated as follows:

$$\mathrm{R} = \sum_{m=1}^{M} \|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2 \tag{10}$$

where $\|\cdot\|_F^2$ denotes the Frobenius norm, and $\|\cdot\|_2^2$ represents the Euclid norm.

Putting Eq. (9) and Eq. (10) into Eq. (3), the loss function of our method can be formulated as follows:

$$\mathrm{L} = \sum_{i,j=1}^{N} \max\{\mathcal{M}_c - y_{ij}(\mathcal{M}_\tau - \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s})), 0\} + \lambda(\sum_{m=1}^{M} \|\mathbf{W}^{(m)}\|_F^2 + \|\mathbf{b}^{(m)}\|_2^2) \tag{11}$$

Using the large adaptive margin algorithm to compute the loss and the back-propagation algorithm to update the deep parameters, the final learned deep ranking model can well distinguish the different individuals by extracting the discriminative feature representations.

3.3. Comparison with two fixed margin loss functions

In order to illustrate how the adaptive margin strategy works, we compare our algorithm with two fixed margin loss functions, namely *the contrastive loss function* and *the triplet loss function*. From Fig. 4, we can conclude that: 1) *The contrastive loss function* has the optimal updating direction of back-propagation, while it penalizes the intra-class distance to zeros and the inter-class distances bigger than a fixed positive margin. 2) *The triplet loss function* minimizes the relative distance between the intra-class samples and the inter-class samples, which also applies the fixed margin strategy. A further critical restriction is that the updating direction of back-propagation deduced by the loss function is not perfect, which can easily make the intra-class distance going up in further iterations. Our *adaptive margin loss function* aims to penalize an adaptive up-margin between the intra-class samples and an adaptive down-margin between the inter-class samples, which can overcome the drawbacks of *the contrastive loss function* in the margin strategy and *the triplet loss function* in both the margin strategy and gradient back-propagation. In most of the deep learning based person re-identification methods, *the triplet loss function* has been proved to be more effective than *the contrastive loss function*, because large number of triplets can be generated in the training process[1]. Benefiting from the adaptive margin strategy, our method can outperform *the triplet loss function* in training the deep CNN. The underlying mechanism is that both the positive distances and negative distances are small at the early iterations, and the positive distances will increase slowly compared to the negative distances with further iterations. Therefore, we will gradually increase the upper-margin M_p to avoid the over-fitting problem by excessively penalizing the positive distances. Meanwhile, we also gradually increase the down-margin M_n to enhance the discrimination of positive pairs from the negative ones. This mechanism can be simply realized by using our adaptive margin strategy, in which we use the average negative distance to model the upper-margin for positive pairs and the average positive distance to model the down-margin for negative pairs. Therefore, the updating of M_p and M_n can jointly consider the variations of negative distance and positive distance, respectively.

[1]Suppose one data set contains M persons from two disjoint camera views, and each person has N images in each camera view, therefore the number of different triplets is $M(M-1)N^3$, while

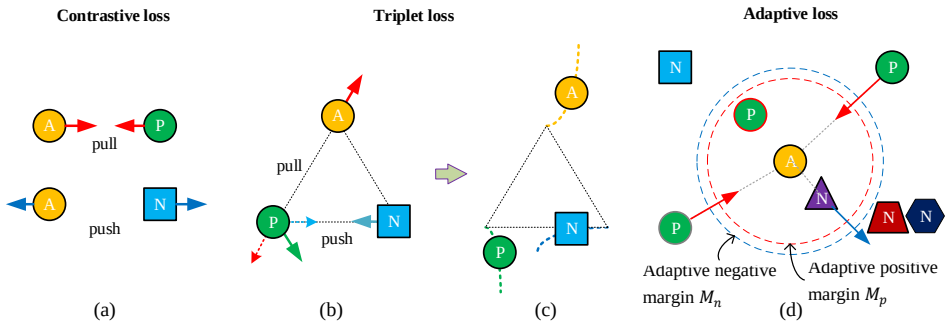

Figure 4: Comparison of our method with two fixed margin approaches, in which (a) shows that the contrastive loss function penalizes the positive distances and negative distances with an optimal gradient back-propagation direction but a fixed margin; (b) and (c) show the triplet loss function minimizes the relative distances between the intra-class samples and inter-class samples with a fixed margin and suboptimal gradient back-propagation direction; (d) shows our adaptive margin loss function penalizes the positive distances and negative distances with an adaptive margin and optimal gradient back-propagation direction.

3.4. Deep neural network

The proposed adaptive margin metric is combined with our part-based deep CNN to implement an end-to-end feature learning and fusion for person re-identification. The whole learning process is illustrated in Fig. 5, in which we firstly organize the training samples into the positive pairs and negative pairs, then we apply the adaptive metric to supervise the leaning of neural network in a Siamese framework. The part-based deep neural network is consisted of three parts, which is small in size but effective in performance. In the following paragraphs, we will introduce each part of the deep neural network in detail.

Global Feature Learning The first part of our network is used for global feature learning, which consists a convolutional layer and max pooling layer. They are used to extract the low-level features of the input images, so as to provide multi-level feature representations to be discriminately learned in the following local sub-network. The input images are in size of $230 \times 80 \times 3$, and are firstly passed through 64 learned filters of size $7 \times 7 \times 3$. Then, the resulting feature maps are passed through a max pooling kernel of size $3 \times 3 \times 3$ with stride 3. Finally, these feature maps are passed through a rectified linear unit (ReLU).

Local Feature Learning The second part of our network is used for local feature learning, which consists of four teams of convolutional layers and max pooling

the number of different positive pairs and negative pairs is $MN^2 + M(M-1)N^2$.

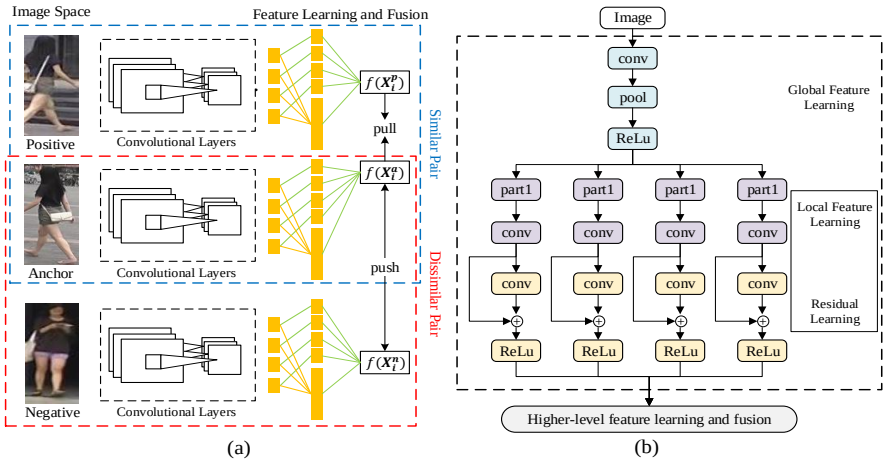

Figure 5: The deep feature learning and fusion neural network. Firstly, we organize the input images into the positive pairs and negative pairs. Secondly, we apply the global feature learning, local feature learning and residual learning in the convolutional layers. Thirdly, we implement the local feature learning and fusion in the fully connected layers. Finally, the concatenated feature representations are fed into the adaptive margin metric for similarity comparison.

layers. We firstly divide the input feature maps into four equal horizontal patches across the height channel, which introduces 4×64 local feature maps of different body parts. Then, we pass each local feature maps through two convolutional layers, and both of them have 32 learned filters of size 3×3. Furthermore, the outputs of the first local convolutional layer are summarized with the outputs of the second local convolutional layer using element-wise operation, which we called residual module. Afterwards, the resulting feature maps are passed through max pooling kernels of size 3×3 with stride 1. Finally, we add a ReLU after each max pooling layer. In order to learn the feature representations of different body parts discriminately, we do not share the parameters among the four paths of convolutional layers.

Feature Learning and Fusion The third part of our network is used for feature learning and fusion sub-network, which consists of four teams of fully connected layers. Firstly, the local feature maps of different body parts are discriminately learned by following two fully connected layers in each path. The dimension of the fully connected layer is 100 and a ReLU is added between them. Then, the discriminately learned local feature representations of the first four fully connected layers are concatenated to be summarized by adding another fully connected layer, whose dimension is 400. Finally, the resulting feature representation is further concatenated with the outputs of the second four fully connected layers to gener-

ate 800 dimensional final feature representations. Similarly, we do not share the parameters among the four fully connected layers to perserve the discrimination of feature representations of different body parts.

4. Optimization

We use the gradient back-propagation method to optimize the parameters of the deep CNN, which is carried out in the mini-batch manner. Therefore, we need to calculate the gradients of the loss function with respect to the feature representations at the output layers. For simplicity, we consider the parameters in the network as a whole and define $\mathbf{\Omega}^{(m)} = [\mathbf{W}^{(m)}, \mathbf{b}^{(m)}]$, and $\mathbf{\Omega} = \{\mathbf{\Omega}^{(1)}, \ldots, \mathbf{\Omega}^{(M)}\}$.

In order to employ the back-propagation algorithm to optimize the network parameters, we compute the partial derivative of the loss function as follows:

$$\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}} = \sum_{i,j=1}^{N} \frac{\partial \mathrm{L}}{\partial f(\mathbf{x}_A^{i,k})} \cdot \frac{\partial f(\mathbf{x}_A^{i,k})}{\partial \mathbf{\Omega}} + \frac{\partial \mathrm{L}}{\partial f(\mathbf{x}_B^{j,s})} \cdot \frac{\partial f(\mathbf{x}_B^{j,s})}{\partial \mathbf{\Omega}} + 2\lambda \sum_{k=1}^{M} \mathbf{\Omega}^{(k)} \tag{12}$$

where the first two terms represent the gradient of the similarity comparison term, and the third term is the gradient of the regularization term.

For simplicity, we define $\mathcal{R} = \mathcal{M}_c - y_{ij}(\mathcal{M}_\tau - \mathcal{D}(\mathbf{x}_A^{i,k}, \mathbf{x}_B^{j,s}))$, then $\partial \mathrm{L}/\partial f(\mathbf{x}_A^{i,k})$ and $\partial \mathrm{L}/\partial f(\mathbf{x}_B^{j,s})$ can be formulated as follows:

$$\frac{\partial \mathrm{L}}{\partial f(\mathbf{x}_A^{i,k})} = \begin{cases} -y_{i,j}(\mathcal{M}_\tau - 2(f(\mathbf{x}_A^{i,k}) - f(\mathbf{x}_B^{j,s}))), \ if \ \ \mathcal{R} > 0; \\ 0, \qquad\qquad\qquad else. \end{cases} \tag{13}$$

$$\frac{\partial \mathrm{L}}{\partial f(\mathbf{x}_B^{j,s})} = \begin{cases} -y_{i,j}(\mathcal{M}_\tau + 2(f(\mathbf{x}_A^{i,k}) - f(\mathbf{x}_B^{j,s}))), \ if \ \ \mathcal{R} > 0; \\ 0, \qquad\qquad\qquad else. \end{cases} \tag{14}$$

From the above derivations, it is clear that the gradients of our adaptive margin metric can be easily calculated given the values of $f(\mathbf{x}_A^{i,k})$, $f(\mathbf{x}_B^{j,s})$ and $\partial f(\mathbf{x}_A^{i,k})/\partial \mathbf{\Omega}$, $\partial f(\mathbf{x}_A^{i,k})/\partial \mathbf{\Omega}$ in each mini-batch, in which they can be obtained by separately running the forward and backward propagation for all the positive pairs and negative pairs. Since the algorithm needs to go through all the pairwise units to accumulate the gradients in each iteration, we call it the adaptive margin gradient descent algorithm. We show the overall process in Algorithm 1.

5. Mini-batch Generation

As the person re-identification datasets usually contains hundreds of or even thousands of pedestrians, it is impossible to load all the positive pairs and negative

Algorithm 1 The adaptive margin gradient descent algorithm

Input: Training samples $\mathbf{X}$, learning rate ω, maximum iterations H, adaptive margin parameters μ and γ, weight parameter λ, updating rate η.
Output: The deep parameters $\mathbf{\Omega}$.
while $h < H$ **do**
    Calculate the output feature representations of $f(\mathbf{x}_A^{i,a})$ and $f(\mathbf{x}_B^{j,s})$ for all the positive pairs and negative pairs in a mini-batch by forward propagation;
    Compute the adaptive margin $\mathcal{M}_p$ and $\mathcal{M}_n$ according to Eq. (6);
    Compute the $\partial f(\mathbf{x}_A^{i,k})/\partial \mathbf{\Omega}, \partial f(\mathbf{x}_A^{i,k})/\partial \mathbf{\Omega}$ according to Eq. (13) and Eq. (14);
    Compute the $\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}}$ according to Eq. (12);
    Update the deep parameters $\mathbf{\Omega}_{h+1} = \mathbf{\Omega}_h - \tau_h \frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}_h}$ and $h \leftarrow h + 1$.
end while

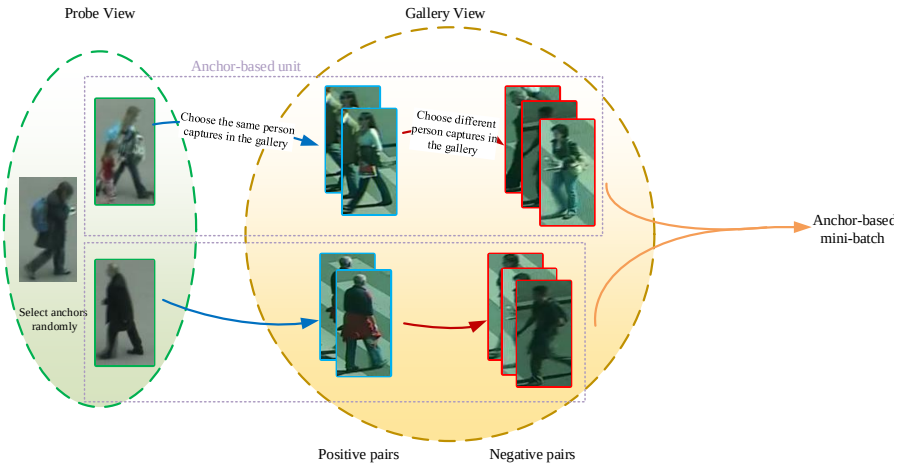

Figure 6: Illustration of the anchor-based minibatch generation. Firstly, we randomly find a subset of the captures of persons in the probe view, and each image is called anchor. Secondly, we randomly find M corresponding positives and N negatives in the gallery, respectively. Finally, the generated positive pairs and negative pairs are fed into the neural network in a min-batch manner.

pairs due to the memory limit and computational speed. Therefore, it is necessary to train the neural network in the mini-batch mode. To be more specific, we only use a small set of anchor-based units to train the neural network in each iteration.

The basic pipeline of anchor-based mini-batch generation is shown in Fig. 6. Given a labeled training set, we randomly find a subset of images in the probe view, where each image is called an *anchor*. Then, we randomly find M corresponding positives and N negatives for each anchor in the gallery view. Finally, the gener-

Algorithm 2 Learning features of the training set in the mini-batch mode

Input:

Labeled training images$\{I_i\}$.

Output:

The network parameters $\mathbf{\Omega}$.

while $n < N$ **do**
    Randomly select A anchors from the training set;
    for all anchors **do**
        Randomly select M positives and N negatives and form the anchor-based unit.
    end for
    Form the mini-batch and $n \leftarrow n + 1$.
end while
for all mini-batches **do**
    Update $\mathbf{\Omega}$ according to Algorithm 1;
end for

ated positive pairs and negative pairs are fed into the neural network in a minbatch manner. By repeating the steps above several times, we can obtain a large number of possible anchor-based units without overburdening the hardware, because random selection mechanism guarantees huge possibilities of the input combination. Algorithm 2 shows the complete batch training process.

6. Experiments

6.1. Datasets and Settings

We evaluate our method on four widely used benchmark datasets, namely PRID2011 [44], Market1501 [45], CUHK01 [26] and 3DPeS [46]. Each of them has at least 2 image for each person from each camera view.

PRID2011: The dataset includes 749 persons, captured by two disjoint cameras, with sequences lengths of 5 to 675 frames. Following the protocol used in [47], we only consider the first 200 persons, who appear in both cameras.

Market1501: The dataset contains 32668 images of 1501 individuals. Each individual is captured by six cameras at most, and two cameras at least. We use the provided fixed training and test set, under both the single-query and multi-query evaluation settings as in [6].

CUHK01: The dataset contains 971 persons captured from two camera views in a campus environment, and there are two images for each person from every

camera view. We utilize the same protocol with [43], where 871 person images are used for training and the rest for testing.

3DPeS: The dataset has 1011 images of 192 persons captured from 8 outdoor cameras with significantly different viewpoints. The image number for each person varies from 2 to 26. We utilize the same protocol as in [48], where half of the persons are used for training and the rest for testing.

Parameter setting The weights are initialized from two zero-mean Gaussian distribution with the standard deviations from 0.01 to 0.001, respectively. The bias terms are set to 0. The learning rate $\omega = 0.01$, the updating rate $\eta = 0.001$, the weight parameter $\alpha = 0.01$.

Evaluation protocol The dataset is separated into the training set and the test set, in which images of same person appear in both sets. The testing set is further divided into probe set and gallery set, and the two sets contains different images of the same person. The result is evaluated by cumulative matching characteristic (CMC) curve [49], which is an estimation of finding the corrected match in the top n match. Final performance is averaged over ten random repeats of the process. Specially for Market1501, we also report mAP result.

6.2. Results

Comprehensive evaluation, including single-query and multi-query, is conducted on the proposed method and compared with the state-of-the-art methods. Specifically, for dataset CUHK01 and 3DPeS, CMC curves are single-shot results; for dataset PRID2011 and Market1501, the Cumulative Matching Characteristics (CMC) curves are multi-shot results.

PRID2011: This dataset is specially designed for video-based person re-identification problem. To make fair comparison, we choose to avoid using any video-based information in our experiments. Results in Table. 1 and Fig. 7(a) show that the proposed method has the best performance in all top-1 to top-50 identification result. Compared with the previous best method discriminative null space method (DNS) [6], our method outperforms it by a large margin (57.8% vs 40.9%) in top-1 accuracy.

Market1501: As a newly proposed large scale person re-identification dataset, the best performance was obtained by DNS [6]. As illustrated in Table. 2 and Fig. 7(b), the proposed method improves the top-1 identification rate from 67.96% to 80.61%, and mAP from 41.89% to 47.63%.

CUHK01: From the results in Table 3 and Fig. 7(c), the previous best top-1 accuracy was achieved by Ejaz *et al.* [14] at 65.0%. Our method obtains 71.90% top-1 accuracy, outperforming all other methods. Besides, top-1 to top-20 identification rates outperform the state-of-the-art-method by a large margin.

Table 1: Identification rates (%) of different methods on PRID2011 dataset

Method	$r = 1$	$r = 10$	$r = 20$	$r = 50$
DeepM [50]	17.9	45.9	55.4	74.1
KISSME [51]	15.0	39.0	52.0	68.0
Mahalanobis [44]	16.0	41.0	51.0	64.0
EIML [52]	16.0	39.0	51.0	68.0
ITML [13]	12.0	36.0	47.0	64.0
LMNN [12]	10.0	30.0	42.0	59.0
DNS [6]	40.9	73.2	81.0	90.4
Ours	**57.8**	**90.3**	**97.7**	**100**

Table 2: Top-1 identification rate and mAP of various methods on Market 1501 dataset

Method	$r = 1$	mAP
BoW [45]	34.38	14.10
Baseline(+Mahalanobis) [45]	36.79	15.08
Baseline(+KISSME[53])	42.70	19.55
SCSP [48]	51.90	26.35
XQDA(LOMO) [28]	54.13	28.41
S-LSTM [54]	61.60	35.31
DNS(LOMO) [6]	67.96	41.89
Ours	**80.61**	**47.63**

Table 3: Identification rates (%) of different methods on CUHK01 dataset

Method	$r = 1$	$r = 5$	$r = 10$	$r = 20$
Ejaz's [14]	65.0	87.0	94.0	97.0
FPNN [40]	20.7	51.0	67.0	82.5
KISSME [53]	14.2	36.0	51.5	69.0
LDM [55]	13.5	36.0	51.4	70.5
RANK [56]	10.4	30.0	46.0	62.0
eSDC [22]	8.8	24.0	36.0	52.8
LMNN [12]	7.3	19.8	30.5	48.0
ITML [13]	5.5	23.5	35.5	50.8
Ours	**71.9**	**91.8**	**95.8**	**97.2**

3DPeS: The proposed method obtains 58.3% top-1 accuracy, outperforming all the other methods, as presented in Table 4 and Fig. 7(d). However, kLFDA [10] achieves the highest top-5 identification rate and ME [48] achieves the highest top-20 identification rate. Yet it is worth noting that the difference of these rates between the state-of-the-art method and our method is relatively small.

Table 4: Identification rates (%) of different methods on 3DPeS dataset

Method	$r = 1$	$r = 5$	$r = 20$
LF [29]	45.5	69.2	89.06
ME [57]	53.3	76.8	**92.78**
kLFDA [10]	54.0	**77.7**	92.3
SCSP [48]	57.3	78.9	91.5
Ours	**58.3**	74.0	88.5

6.3. Analysis

As observed in our experiments, hyper parameters μ, γ influences the performance of the proposed method. We give an empirical analysis in PRID2011 dataset in Section 6.3.1. We also analyze the influence of different loss functions and residual learning module, as well as the ranking results.

6.3.1. Parameter Analysis

In Section 3.1, we introduce Eq. (6) as a nonlinear mapping method for calculating the average distance. As shown in Fig. 3, hyper parameters μ, γ directly control the dynamic margin. We analyze the influence by changing one parameter while fixing the other one.

The influence of μ on $\mathcal{M}_p$. As depicted in Fig. 3, $\mathcal{M}_p$ in Eq. (6) increases when μ decreases (from red, green to blue line). If μ is too huge, $\mathcal{M}_p$ can get extremely small, leading to overfitting. Similarly, too small μ gives huge down margin, making the constraint too weak and leads to underfitting. The Fig. 8(b) plots the CMC curves with a fixed γ and varying μ in PRID2011 dataset. The performance is optimal when $\mu = 8$.

The influence of γ on $\mathcal{M}_n$. When γ increases (from blue, green to red line), $\mathcal{M}_n$ diminishes, as depicted in Fig. 3. If γ is too huge, underfitting might happen, due to small down-margin $\mathcal{M}_n$. Tiny γ will lead to huge down-margin and overfitting. Results on PRID2011 with a fixed μ and varying γ in Fig. 8(d) shows that, the performance reaches best when $\gamma = 2.1$.

We can conclude that, although the identification rate will stay robust to a slight change in hyper parameters, tuning the 2 hyper parameters is necessary to avoid overfitting or underfitting, and to get optimal perforamance.

6.3.2. Analysis of loss functions and the residual module

Overall, the proposed method mainly consists of two novel parts: i) the residual learning module in the proposed deep network; ii) the adaptive contrastive loss function. To review how each part contributes to the performance improvement, we implement five variants of the proposed person re-identification method, and compare their functions in the literature:

Variant 1: We replace the residual learning module with a convolution layer and use the conventional contrastive loss function to train the network. We refer this network as PCNN-C;

Variant 2: We replace the residual learning module with a convolution layer but retain the adaptive contrastive loss function. We refer this network as PCNN-AC;

Variant 3: We retain the residual learning module (as in Section 3.4). However, conventional contrastive loss function is applied. We refer this network as RCNN-C;

Variant 4: We retain the residual learning module (as in Section 3.4). However, conventional triplet loss function is applied. We refer this network as RCNN-T;

Variant 5: We retain the residual learning module as in Section 3.4 and adaptive contrastive loss function. We refer this network as RCNN-AC. This is our proposed integrative method.

The comparison results are shown in Fig. 9 and Table. 5. The following discussion mainly focus on the influence of two modules: adaptive contrastive loss function and residual learning module.

Table 5: Comparison of module influence in terms of identification rates (%) on different datasets

Dataset	method	$r = 1$	Dataset	method	$r = 1$
PRID2011	PCNN-C	48.5	Market1501	PCNN-C	72.3
	PCNN-AC	52.4		PCNN-AC	78.7
	RCNN-C	53.8		RCNN-C	74.4
	RCNN-T	55.2		RCNN-T	76.6
	RCNN-AC	**57.5**		**RCNN-AC**	**80.6**
CUHK01	PCNN-C	60.8	3DPeS	PCNN-C	41.7
	PCNN-AC	70.6		PCNN-AC	52.8
	RCNN-C	62.6		RCNN-C	46.7
	RCNN-T	65.0		RCNN-T	48.2
	RCNN-AC	**71.9**		**RCNN-AC**	**58.3**

The influence of adaptive contrastive loss function. We compare the adaptive margin loss function with the conventional contrastive loss function and the triplet loss function, respectively. Compared with contrastive loss function, the top-1 identification rate of PRID2011, Market1501,CUHK01 and 3DPeS increases by 3.7%, 6.2%, 9.3%, 11.6% with residual model, and 3.9%, 6.4%, 9.8%, 11.1%

without residual model, respectively. Compared with the triplet loss, the top-1 identification rate of PRID2011, Market1501,CUHK01 and 3DPeS increases by 2.3%, 4.0%, 6.9%, 10.1%, respectively.

The influence of residual learning module. On adopting residual learning module, the top-1 identification rate of PRID2011, Market1501,CUHK01 and 3DPeS increases by 5.2%, 2.0%, 1.6%, 5.3% on average, respectively. Besides, the overall identification rate also improves on applying the residual learning module on the 4 datasets.

The influence of combining 2 modules. On applying both the adaptive contrastive loss function and residual learning module, compared with conventional contrastive loss function, the top-1 identification rate of PRID2011, Market1501, CUHK01 and 3DPeS increases by 9.0%, 8.3%, 11.1%, 16.6% on average, respectively.

We can reach the following 3 empirical conclusion from the experimental results: 1) Adaptive margin loss function performs better than conventional contrastive and triplet loss function; 2) Residual module contributes to the improvement of the method's performance; 3) Combining adaptive margin loss function and the proposed deep architecture achieves higher performance.

6.3.3. Ranking result analysis

Some representative top-5 ranking results on the 4 datasets are shown in Fig. 10. Most of the ranking results have high similarity, and some of them are even hard to discriminate by human eyes. According to [58], if we can emphasize more on narrowing the distance between the difficult negative references and its anchor, the accuracy of the ranking result may be enhanced further. In the future work, incorporating the mechanism of hard-negative strategy to our adaptive contrastive loss function may also be expected to improve the performance of the proposed method.

7. Conclusion

In this paper, we present a novel method for person re-identification, which applies an adaptive margin strategy to learn a deep ranking model in the siamese framework. In order to learn discriminative and stable feature representations, we build a part-based deep neural network, in which different body parts are first discriminately learned in the global sub-network and local sub-network, and then fused in the fusion sub-network. The output features are further fed into the loss layer to penalize an adaptive margin between the positive pairs and negative pairs, so as to learn the deep model in an incremental manner. Experimental results on

the PRID2011, Market1501, CUHK01 and 3DPeS datasets show that our method outperforms the state-of-the-art in person re-identification.

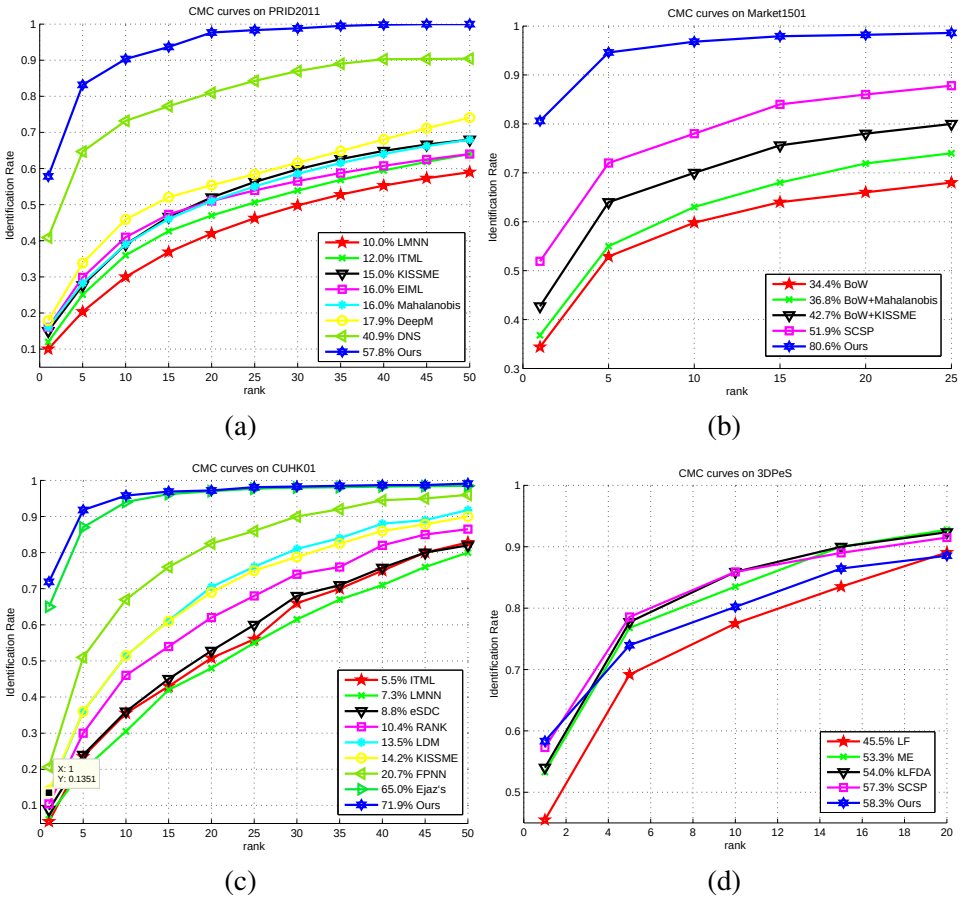

Figure 7: CMC curves for method comparison on: (a) the PRID2011 dataset (multi-query, 100 gallery identities); (b) the Market1501 dataset (multi-query, 722 gallery identities); (c) the CUHK01 dataset (single-query, 100 gallery identities); (d) the 3DPeS dataset (single-query, 96 gallery identities). The top-1 identification rates are shown with the method names.

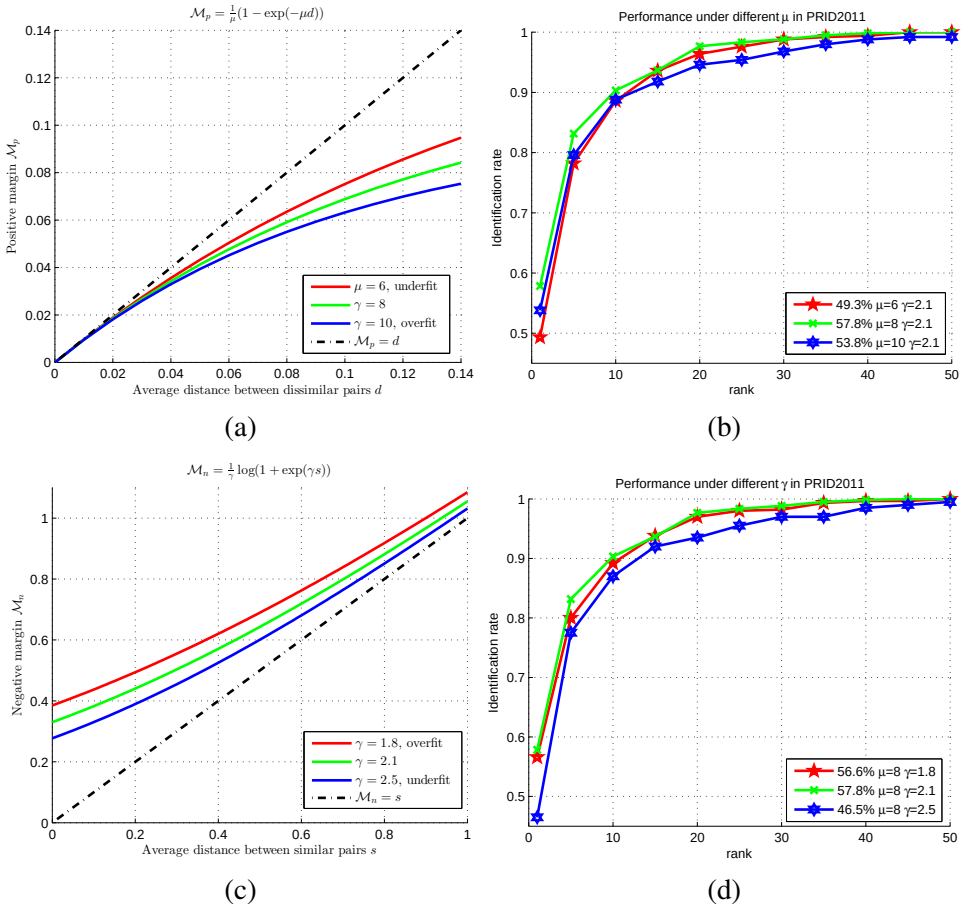

Figure 8: The analysis of parameter on PRID2011 dataset. (a) illustration of the influence of parameter μ: when $\mu = 6, \mathcal{M}_p$ is rather substantial, the intra-class distance cannot be effectively narrowed, therefore leading to underfitting; when $\mu = 10, \mathcal{M}_p$ is rather small, the intra-class distance is condensed too much, therefore leading to overfitting; thus, choosing μ between 6 and 10 is ideal; (b) CMC curves with a fixed γ and varying μ, the deep architecture performs best when $\mu = 8$; (c) illustration of the influence of parameter γ: when $\gamma = 1.8, \mathcal{M}_n$ is rather big, the inter-class distance are driven to be too huge, therefore leading to overfitting; when $\gamma = 2.5, \mathcal{M}_p$ is rather small, the inter-class distance cannot be effectively enlarged, therefore leading to underfitting; thus, choosing γ between 1.8 and 2.5 is ideal; (d) CMC curves with a fixed μ and varying γ, the deep architecture performs best when $\gamma = 2.1$.

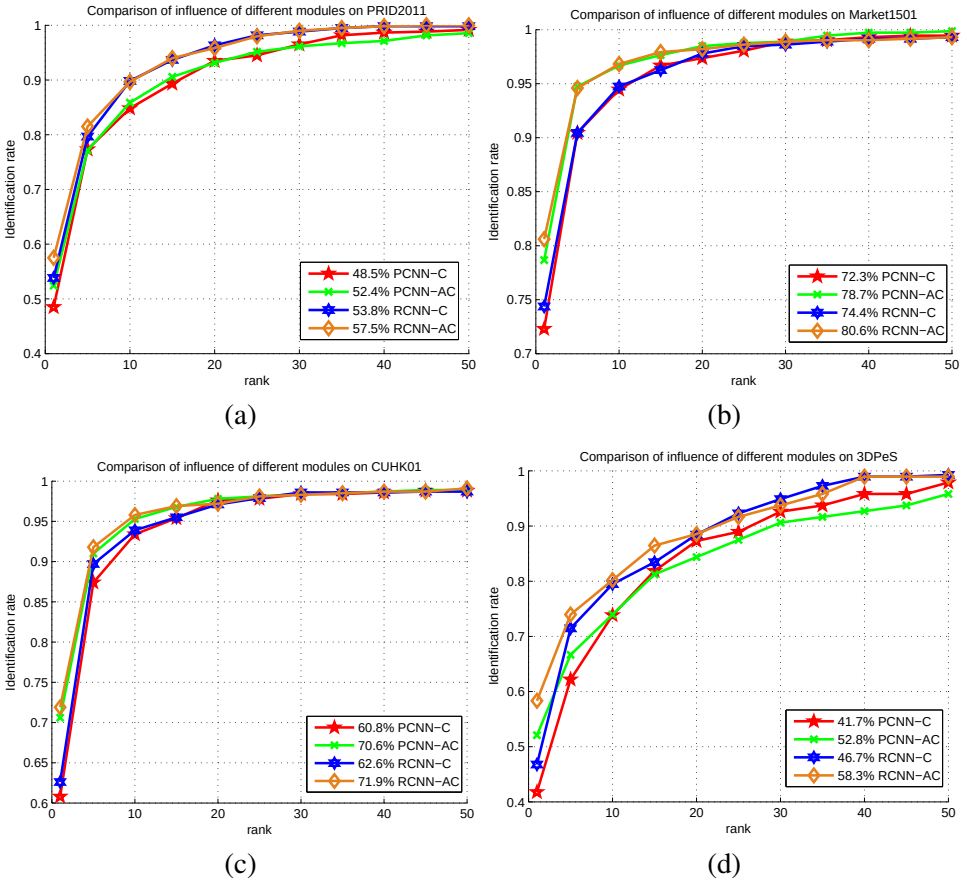

Figure 9: The performance (CMC curves) comparison of module function on: (a) the PRID2011 dataset; (b) the Market1501 dataset; (c) the CUHK01 dataset; (d) the 3DPeS dataset. The top-1 identification rates are shown after the method names. PCNN-C replaces the residual learning module with a convolution layer and adopts conventional contrastive loss function; PCNN-AC replaces the residual learning module with a convolution layer and adopts adaptive contrastive loss function; RCNN-C adopts residual learning module and conventional contrastive loss function; RCNN-AC adopts residual learning module and adpative contrastive loss function.

Figure 10: Ranking result of PRID2011, Market1501, CUHK01 and 3dPeS. The left row is a query in test phase. The right block in green frame corresponds to the returned top5 results of the query. The one marked by red is the matched reference of the query.

References

References

[1] S. Zhang, J. Wang, Z. Wang, Y. Gong, Y. Liu, Multi-target tracking by learning local-to-global trajectory models, Pattern Recognition 48 (2) (2015) 580–590.

[2] F. Solera, S. Calderara, R. Cucchiara, Socially constrained structural learning for groups detection in crowd, IEEE transactions on pattern analysis and machine intelligence 38 (5) (2016) 995–1008.

[3] W. Hu, T. Tan, L. Wang, S. Maybank, A survey on visual surveillance of object motion and behaviors, IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) 34 (3) (2004) 334–352.

[4] R. R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision, Springer, 2016, pp. 791–808.

[5] S. Zhou, J. Wang, Q. Hou, Y. Gong, Deep ranking model for person re-identification with pairwise similarity comparison, in: Pacific Rim Conference on Multimedia, Springer, 2016, pp. 84–94.

[6] L. Zhang, T. Xiang, S. Gong, Learning a discriminative null space for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[7] N. McLaughlin, J. Martinez del Rincon, P. Miller, Recurrent convolutional network for video-based person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[8] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European conference on computer vision, Springer, 2008, pp. 262–275.

[9] R. Zhao, W. Ouyang, X. Wang, Learning mid-level filters for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 144–151.

[10] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision, Springer, 2014, pp. 1–16.

[11] Z. Li, S. Chang, F. Liang, T. S. Huang, L. Cao, J. R. Smith, Learning locally-adaptive decision functions for person verification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3610–3617.

[12] K. Q. Weinberger, J. Blitzer, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, in: Advances in neural information processing systems, 2005, pp. 1473–1480.

[13] J. V. Davis, B. Kulis, P. Jain, S. Sra, I. S. Dhillon, Information-theoretic metric learning, in: Proceedings of the 24th international conference on Machine learning, ACM, 2007, pp. 209–216.

[14] E. Ahmed, M. Jones, T. K. Marks, An improved deep learning architecture for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[15] S. Ding, L. Lin, G. Wang, H. Chao, Deep feature learning with relative distance comparison for person re-identification, Pattern Recognition 48 (10) (2015) 2993–3003.

[16] Y. Wang, C. Xu, S. You, D. Tao, C. Xu, Cnnpack: Packing convolutional neural networks in the frequency domain, in: Advances in Neural Information Processing Systems, 2016.

[17] I. Hubara, M. Courbariaux, D. Soudry, R. El-Yaniv, Y. Bengio, Binarized neural networks, in: Advances in Neural Information Processing Systems, 2016, pp. 4107–4115.

[18] C. Xu, D. Tao, C. Xu, Y. Rui, Large-margin weakly supervised dimensionality reduction, in: Proceedings of the 31st international conference on machine learning (ICML-14), 2014, pp. 865–873.

[19] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[20] B. Ma, Y. Su, F. Jurie, Bicov: a novel image representation for person re-identification and face verification, in: British Machive Vision Conference, 2012, pp. 11–pages.

[21] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE, 2010, pp. 2360–2367.

[22] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3586–3593.

[23] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, V. Murino, Custom pictorial structures for re-identification., in: BMVC, Vol. 1, Citeseer, 2011, p. 6.

[24] Z. Wu, Y. Li, R. J. Radke, Viewpoint invariant human re-identification in camera networks using pose priors and subject-discriminative features, Pattern Analysis and Machine Intelligence, IEEE Transactions on 37 (5) (2015) 1095–1108.

[25] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2288–2295.

[26] W. Li, X. Wang, Locally aligned feature transforms across views, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3594–3601.

[27] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (3) (2013) 653–668.

[28] A. Mignon, F. Jurie, Pcca: A new approach for distance learning from sparse pairwise constraints, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2666–2672.

[29] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 3318–3325.

[30] H. V. Nguyen, L. Bai, Cosine similarity metric learning for face verification, in: Computer Vision–ACCV 2010, Springer, 2011, pp. 709–720.

[31] C. C. Loy, C. Liu, S. Gong, Person re-identification by manifold ranking, in: Image Processing (ICIP), 2013 20th IEEE International Conference on, IEEE, 2013, pp. 3567–3571.

[32] D. Chen, Z. Yuan, G. Hua, N. Zheng, J. Wang, Similarity learning on an explicit polynomial kernel feature map for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1565–1573.

[33] M. Hirzer, P. M. Roth, M. Köstinger, H. Bischof, Relaxed pairwise learned metric for person re-identification, in: Computer Vision–ECCV 2012, Springer, 2012, pp. 780–793.

[34] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, IEEE transactions on cybernetics.

[35] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, IEEE Transactions on Information Forensics and Security 12 (5) (2017) 1005–1016.

[36] J. Yu, D. Tao, M. Wang, Y. Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE transactions on cybernetics 45 (4) (2015) 767–779.

[37] L. Wu, C. Shen, A. van den Hengel, Deep recurrent convolutional networks for video-based person re-identification: An end-to-end approach, CoRR abs/1606.01609.
URL `http://arxiv.org/abs/1606.01609`

[38] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, CoRR abs/1701.08398.
URL http://arxiv.org/abs/1701.08398

[39] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned CNN embedding for person re-identification, CoRR abs/1611.05666.
URL http://arxiv.org/abs/1611.05666

[40] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[41] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[42] D. Yi, Z. Lei, S. Liao, S. Z. Li, Deep metric learning for person re-identification, in: Pattern Recognition (ICPR), 2014 22nd International Conference on, IEEE, 2014, pp. 34–39.

[43] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

[44] M. Hirzer, C. Beleznai, P. M. Roth, H. Bischof, Person re-identification by descriptive and discriminative classification, in: Scandinavian conference on Image analysis, Springer, 2011, pp. 91–102.

[45] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.

[46] D. Baltieri, R. Vezzani, R. Cucchiara, Sarc3d: a new 3d body model for people tracking and re-identification, in: International Conference on Image Analysis and Processing, Springer, 2011, pp. 197–206.

[47] J. You, A. Wu, X. Li, W.-S. Zheng, Top-push video-based person re-identification, arXiv preprint arXiv:1604.08683.

[48] D. Chen, Z. Yuan, B. Chen, N. Zheng, Similarity learning with spatial constraints for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1268–1277.

[49] D. Gray, S. Brennan, H. Tao, Evaluating appearance models for recognition, reacquisition, and tracking, in: Proc. IEEE International Workshop on Performance Evaluation for Tracking and Surveillance (PETS), Vol. 3, Citeseer, 2007.

[50] D. Yi, Z. Lei, S. Liao, S. Z. Li, Deep metric learning for person re-identification, in: International Conference on Pattern Recognition, 2014, pp. 34–39.

[51] M. Kstinger, M. Hirzer, P. Wohlhart, P. M. Roth, Large scale metric learning from equivalence constraints, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.

[52] M. Hirzer, P. M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 203–208.

[53] S. Khamis, C. H. Kuo, V. K. Singh, V. D. Shet, L. S. Davis, Joint Learning for Attribute-Consistent Person Re-Identification, ECCV Workshop on Visual Surveillance and Reidentification, 2014.

[54] R. R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, European Conference on Computer Vision.

[55] M. Guillaumin, J. Verbeek, C. Schmid, Is that you? metric learning approaches for face identification, in: ICCV, 2009, pp. 498–505.

[56] B. Mcfee, G. R. G. Lanckriet, Metric learning to rank., in: International Conference on Machine Learning, 2010.

[57] S. Paisitkriangkrai, C. Shen, V. D. H. Anton, Learning to rank in person re-identification with metric ensembles, Computer Science (2015) 1846–1855.

[58] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, Computer Vision and Pattern Recognition (2015) 815–823.