Deep Self-Paced Learning for Person Re-Identification

Sanping Zhou[a], Jinjun Wang[a,*], Deyu Meng[b], Xiaomeng Xin[a], Yubing Li[c], Yihong Gong[a], Nanning Zheng[a]

[a]*The institute of artificial intelligence and robotic, Xi'an Jiaotong University, Xian Ning West Road No.28, Shaanxi, 710049, P.R. China*
[b]*School of Mathematics and Statistics, Xi'an Jiaotong University, Xian Ning West Road No.28, Shaanxi, 710049, P.R. China*
[c]*School of the Electronic and Information Engineering, Xi'an Jiaotong University, Xian Ning West Road No.28, Shaanxi, 710049, P.R. China*

Abstract

Person re-identification (Re-ID) usually suffers from noisy samples with background clutter and mutual occlusion, which makes it extremely difficult to distinguish different individuals across the disjoint camera views. In this paper, we propose a novel deep self-paced learning (DSPL) algorithm to alleviate this problem, in which we apply a self-paced constraint and symmetric regularization to help the relative distance metric training the deep neural network, so as to learn the stable and discriminative features for person Re-ID. Firstly, we propose a soft polynomial regularizer term which can derive the adaptive weights to samples based on both the training loss and model age. As a result, the high-confidence fidelity samples will be emphasized and the low-confidence noisy samples will be suppressed at early stage of the whole training process. Such a learning regime is naturally implemented under a self-paced learning (SPL) framework, in which samples weights are adaptively updated based on both model age and sample loss using an alternative optimization method. Secondly, we introduce a symmetric regularizer term to revise the asymmetric gradient back-propagation derived by the relative distance metric, so as to simultaneously minimize the intra-class distance and maximize the inter-class distance in each triplet unit. Finally, we build a part-based deep neural network, in which the features of different body parts are first discriminately learned in the lower convolutional layers and then fused in the higher fully connected layers. Experiments on several benchmark datasets have demonstrated the superior performance of our method as compared with the state-of-the-art approaches.

Keywords: Person Re-identification, Deep Convolutional Neural Network, Self-Paced Learning, Metric Learning.

1. Introduction

Person re-identification (Re-ID) has become an active research topic in the field of computer vision, because of its wide application in the video surveillance community. Given one single shot or multiple shots of a target, person Re-ID concerns the problem of matching the same person among a set of gallery candidates captured from the disjoint camera views [1–4]. It is a very challenging task due to noisy samples with mutual occlusion and background clutter that makes the large appearance variations across different camera views [5, 6]. Therefore, the key to improve the identification performance is to learn the stable and discriminative features for representation.

The fundamental person Re-ID problem is to compare an image of each interested target seen in a probe camera view to a large number of candidates captured from a gallery camera view which has no overlap with the probe one [7]. If a true match to the probe exists in the gallery, it should have a higher similarity score as compared with the incorrect matches. Previous efforts for solving this problem primarily focus on the following two aspects: 1) developing robust feature descriptors to handle the variations in person's appearance, and 2) designing discriminative distance metrics to measure the similarity of person's images. For the first category, different cues are

*Corresponding author: Tel.: +86-029-83395146; Fax: +86-029-83395175;
Email address: sanpingzhou@stu.xjtu.edu.cn (Sanping Zhou)

employed for the stable and discriminative features. Representative descriptors include the Local Binary Pattern (LBP) [8], Ensemble of Local Feature (ELF) [9] and Local Maximal Occurrence (LOMO) [10]. For the second category, labeled images are used to train a distance metric, in which the intra-class distance is minimized while the inter-class distance is maximized. Typical metric learning methods include the Locally Adaptive Decision Function (LADF) [11], Large Margin Nearest Neighbor (LMNN) [12] and Information Theoretic Metric Learning (ITML) [13]. Since both line of works regard the feature extraction and metric learning processes as two disjoint steps, their performances are limited.

In the past two years, the deep convolutional neural network (CNN) based methods [14–19] have been proposed to combine the feature extraction and metric learning into an end-to-end learning framework, in which a neural network is built to extract the stable and discriminative features under the supervision of a suitable distance metric. Benefit from the powerful representation capability of the deep CNN, this line of methods have achieved promising results on the benchmark datasets for person Re-ID. The relative distance metric [20] has been widely used as loss function in the deep learning based methods for visual recognition. Compared with the well-known softmax loss [21], it is a better choice for the zero-shot recognition problem, because of the training set doesn't have the same identity with the testing set. The relative distance metric aims to maximize the relative distance between the positive pair and negative pair in each triplet unit, which can generate a large number of triplet inputs even using a small number of training samples. Therefore, it is very suitable choice for the person Re-ID problem which not only is a zero-shot problem but also can only provide the small-scale dataset for training.

To further improve the identification performance, our observation shows that the following three issues should also be addressed in the learning process. Firstly, the order and weight of training samples should be considered, as shown in Fig. 1, otherwise it might be easy to cause the unstable learning due to the noisy samples or outliers with mutual occlusion and background clutter. Secondly, it is unsuitable to directly apply the distance metric to supervise the training process of deep CNN without any regularization to the gradient back-propagation. Because most of the deep learning tools, such as Caffe [22] and Tensorflow [23], take the gradient back-propagation algorithm to optimize the deep parameters. Thirdly, the neural network should be relatively small and include the part processing module, due to the person Re-ID is a fine-grained problem and the dataset for person Re-ID is usually in small size. As a consequence, it is very urgent to study the three aspects of problems in the training process.

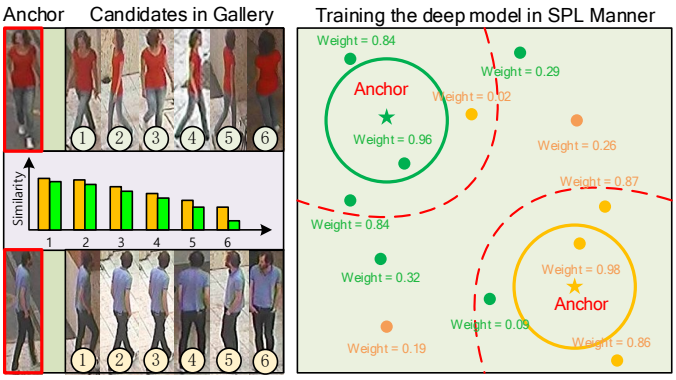

Figure 1: Illustration of our SPL motivations in dealing with the noisy training samples or outliers. The left column shows some typical positive candidates to two anchor images, in which the similarity scores of these positive candidates to the anchor vary from large to small with the incensement of indexes. The right column shows the SPL training strategy , in which the derived weighting scheme will adaptively update the sample weights according to the training loss and model age. Therefore, high-confidence fidelity samples will be emphasized and the the low-confidence noisy samples will be suppressed at early stage of the whole learning process.

In this paper, we propose a novel deep self-paced learning (DSPL) algorithm to adaptively update the weights to samples and regularize the gradient back-propagation of relative distance metric [20] in the learning process, so as to further improve the identification performance of deep neural network for person Re-ID. In order to extract the stable and discriminative features, we firstly build a part-based deep neural network, in which the features of different body parts are discriminately learned in the lower convolutional layers and then fused in the higher fully connected layers. Then, we introduce the self-

paced learning (SPL) theory [24] into the training framework, in which samples can be ranked in a self-paced manner by applying a novel soft polynomial regularizer term to adaptively update the weights according to both the model age and sample loss in each iteration. Specially, the high-confidence fidelity samples will be emphasized and the the low-confidence noisy samples will be suppressed at early stage of the whole learning process. Therefore, the neural network can be trained in a stable process by gradually involving the faithful samples from easy to hard. In addition, a symmetric regularizer term is introduced to overcome the drawback of relative distance metric in gradient back-propagation. As a result, the intra-class distance is minimized and the inter-class distance is maximized by regularizing the asymmetric gradient back-propagation in each triplet unit. Extensive experimental results on several benchmark datasets have shown that our method performs much better than the state-of-the-art approaches.

In summary, the main contributions of this paper can be highlighted as follows:

- We propose a novel DSPL algorithm to supervise the learning of deep neural network, in which a soft polynomial regularizer term is proposed to gradually involve the faithful samples into training process in a self-paced manner.
- We optimize the gradient back-propagation of relative distance metric by introducing a symmetric regularizer term, which can convert the back-propagation from the asymmetric mode to a symmetric one.
- We build an effective part-based deep neural network, in which features of different body parts are first discriminately learned in the lower convolutional layers and then fused in the higher fully connected layers.

The rest of our paper is organized as follows: Section 2 reviews some of the related works. In Section 3, we describe the proposed method, including the DSPL algorithm and deep neural network. The experimental results and corresponding analysis are presented in Section 4. Conclusion comes in Section 5.

2. Related work

In this section, we review two lines of related works, namely the *Person Re-ID* and *Self-Paced Learning*, which are briefly introduced in the following paragraphs.

2.1. Person Re-ID

Extensive works have been reported to address the person Re-ID problem, which mainly focus on several aspects of the issue, such as developing robust feature descriptors, designing distinctive distance metrics and learning stable and discriminative deep features for representation. Blew we will give a brief review of some representative ones.

The feature designing based methods mainly focus on developing robust feature descriptors which are invariant to the view angles, lighting conditions, body poses and background clutters. For example, Zhao et al. [10] learned a mid-level filter from patch cluster to achieve cross-view invariance. In [25], Liao et al. constructed a feature descriptor which analyzed the horizontal occurrence of local features and maximized the occurrence to obtain a robust feature representation against viewpoint changes. Ma et al. [26] presented the person image via covariance descriptor which was robust to illumination changes and background variations. In [27], Farenzena et al. augmented maximally stable color regions with histograms for person representation. Zhao et al. [28] learned the distinct salience features to distinguish the matched person from others. In [29], Chen et al. employed a pre-learned pictorial structure model to localize the body parts more accurately. Wu et al. [30] introduced a viewpoint invariant descriptor, which took the viewpoint of the human into account by using what they called a pose prior learned from the training data. In [31], Kviatkovsky et al. investigated the intra-distribution structure of color descriptor, which was invariant under certain illumination changes. Li et al. [32] matched person images observed in different camera views with complex cross-view transformations and applied it to the person Re-ID problem. These methods aim to improve the person Re-ID performance by developing a fixed

feature descriptor, however the adaptive feature learning is not addressed.

The metric learning based methods aim to find a mapping function from the feature space to another distance space where feature vectors from the same person are more similar than those from different ones. For example, Zheng et al. [33] proposed a relative distance learning method from the probabilistic prospective. In [34], Mignon et al. learned a distance metric from the sparse pairwise similarity constraints. Pedagadi et al. [35] utilized the LADF to map high dimensional features into a more discriminative low dimensional space. In [8], Xiong et al. further extended the LADF and several other metric learning methods by using kernel tricks and different regularizers. Nguyen et al. [36] measured the similarity of face pairs through the cosine similarity, which is closely related to the inner product similarity. In [37], Loy et al. casted the person Re-ID problem as an image retrieval task by considering the listwise similarity. Chen et al. [38] proposed a kernel based metric learning method to explore the nonlinearity relationship of samples in the feature space. In [39], Hirzer et al. learned a discriminative distance metric by using the relaxed pairwise constraints. Prosser et al. [40] developed a ranking model using a support vector machine. These methods learn a specific distance metric mainly based on feature representation extracted by several fixed feature descriptors, which may influence the performance of metric learning.

Different from the above mentioned two lines of methods, the deep learning based methods usually incorporate the feature extraction and metric learning into an end-to-end learning framework, in which a deep neural network is built to extract features from the input images and a distance metric is used to compute the loss and back-propagate the gradients. For example, Ahmed et al. [14] proposed a novel deep neural network which took the pairwise images as inputs, and outputted a similarity score indicating whether the two input images were the same person or not. In [41], Xiao et al. applied a domain guided dropout algorithm to improve the performance of deep CNN in extracting general feature representations. Ding et al. [15] introduced a triplet neural network to learn the relative similarity under supervision of the triplet loss. In [42], Wang et al. proposed a unified triplet loss and siamese deep architecture, which can jointly extract single-image and cross-image feature representations. Zhang et al. [43] incorporated the deep hash learning into a triplet formulation and efficiently improved the identification speed. In [44], Yi et al. constructed a siamese architecture to learn pairwise similarity and used body part strategy to design the neural network. Li et al. [45] proposed a novel filter pairing neural network to model body part displacements by using the patch matching layers to match the filter responses of local patches. In [46], Chen et al. learned a view-specific feature transformation by considering the camera correlation in the deep learning framework. Yan et al. [47] proposed a progressive fusion framework based on the LSTM, so as to aggregate the frame-wise human region representation and yield a sequence level feature representation for person Re-ID. These methods usually incorporate the feature extraction and metric learning into a joint framework mainly based on the general neural networks, such as AlexNet [21] and VGGNet [48], without applying an effective part strategy in the neural networks, which may be inappropriate for the person Re-ID problem.

2.2. Self-Paced Learning

The SPL theory is inspired by the cognitive process of human beings, where samples are involved into the training process from easy to hard ones [49]. As an effective strategy to suppress the side effects of noisy samples or outliers, the SPL based methods have been witnessed the great successes in various machine learning fields [24]. For example, Jiang et al. [50] incorporated the diversity concept into a SPL framework to deal with the event detection and action recognition problem. In [51], Zhang et al. integrated the multiple-instance learning problem into a SPL regime, so as to improve the performance in co-saliency detection. Lee and Grauman [52] proposed a self-paced approach to gradually learn from the complex samples in visual category discovery. In [53], Supancic and Ramanan applied the SPL theory to choose an appropriate framework

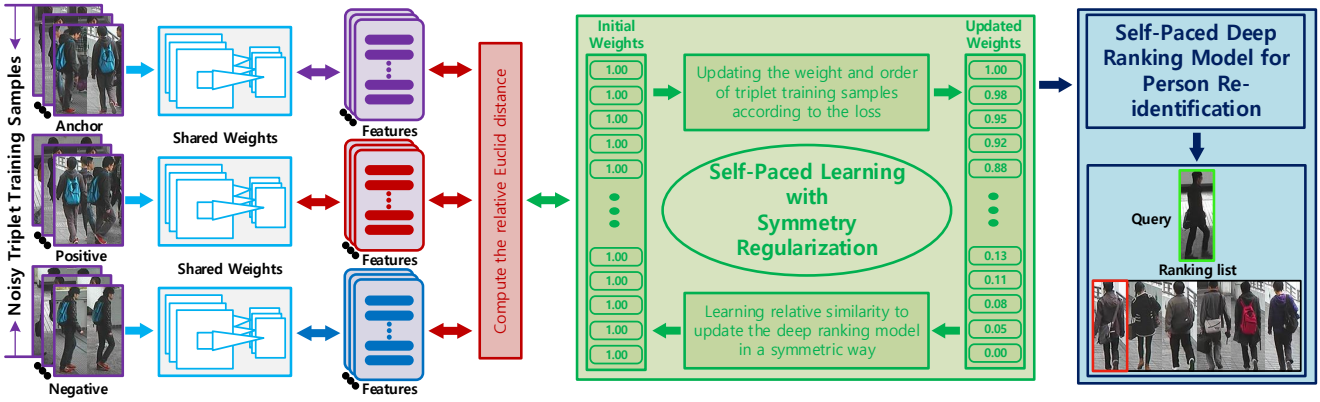

Figure 2: The framework of our deep self-paced person re-identification method. Take the rare images as inputs, our method can effectively alleviate the side effects of noisy training samples or outliers by imposing adaptive weights on them in the relative similarity comparison framework. The SPL constraint gives smaller weights to the noisy low-confidence samples and larger weights to the clean high-confidence samples. As a result, the generalization ability of neural network can be gradually strengthened to deal with the cross-view appearance variations.

to learn good appearance model for long-term tracking. Tang et al. [54] proposed a self-paced domain adaptation method to adapt a object detector from the image domain to the video domain. In [55], Li et al. proposed a multi-objective method to enhance the convergence of the SPL based algorithms. Zhao et al. [56] proposed a novel matrix factorization learning methodology by introducing a soft self-paced regularizer term to impose adaptive weights to samples. In [57], Liang et al. proposed a self-paced cross modal subspace matching method which can gradually chooses the faithful samples to train the model by updating weights in a self-paced manner. Lin et al. [58] developed a novel cost-effective framework to deal with the face identification problem, which utilized the high-confidence and low-confidence samples in both the self-paced and active user-query way. These methods mainly apply the SPL theory in the traditional metric learning framework, and we do not see its application in the popular deep learning framework.

3. The proposed method

Let $\mathbf{X} = \{\mathbf{X}_i | (\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)\}_{i=1}^N$ be the triplet training units, where $\{\mathbf{x}_i^a, \mathbf{x}_i^p\}$ denotes the positive pair, $\{\mathbf{x}_i^a, \mathbf{x}_i^n\}$ represents the negative pair and N denotes the number of total triplet units. The goal of our deep architecture is to learn the filter weights and biases that minimizes the ranking error from the output layer. A recursive function for an K-layer deep model can be formulated as follows:

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

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

3.1. Deep self-paced person Re-ID

The idea of our method is shown in Fig 2, in which we aim to learn a deep ranking model by using the relative similarity comparison in a self-paced manner. The deep ranking model learned in a self-paced manner can be formulated as follows:

$$\mathrm{L} = \sum_{i=1}^{N} u_i \mathcal{R}(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) + \mathcal{G}(\mathbf{u}, \lambda, \vartheta) + \zeta \mathcal{S}(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) + \xi \mathcal{P}(\mathbf{W}, \mathbf{b}), \tag{2}$$

where $\mathbf{u} = [u_1, \ldots, u_N]^T$ are the weights of all samples, λ, ϑ are the model age parameters, ζ, ξ are the weights of regularizer term. Our method can jointly pull the positive pairs and push the negative pairs in each triplet unit. Specially, the relative similarity term $\mathcal{R}(\cdot)$ maximizes the relative distances between

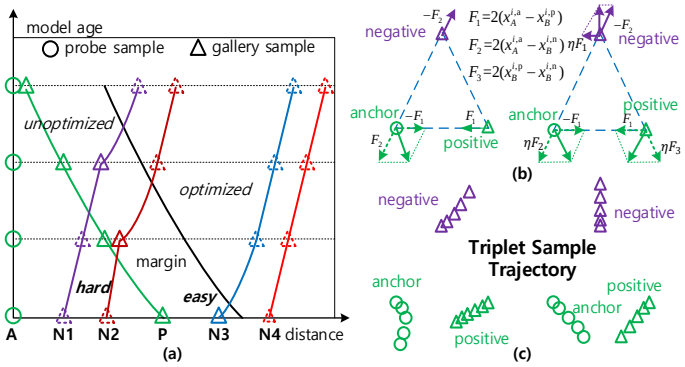

Figure 3: Illustration of the self-paced learning strategy and symmetric gradient back-propagation constraint used in the training process. The left figure shows how the triplet units are gradually involved into the training process with the model age going on, in which the solid triangles denotes the current chosen training samples, and the dotted triangles represent the uninvolved or optimized training samples. The top right figure shows the gradient flow derived by the conventional triplet formulation and the one constrained by the symmetric regularizer term. The bottom right shows the two corresponding motion trajectories driven by the resulting gradient flows, in which our method can simultaneously minimize the intra-class distance and maximize the inter-class distance in each triplet.

the positive pairs and negative pairs, the self-paced regularizer term $\mathcal{G}(\cdot)$ updates the sample weights in a self-paced manner, the symmetric regularizer term $\mathcal{S}(\cdot)$ revises the gradient back-propagation in a symmetric way, and the parameter regularizer term $\mathcal{P}(\cdot)$ smoothes the parameter of the deep CNN. In the following paragraphs, we explain these terms in detail.

Relative similarity term The relative similarity comparison metric has been widely applied in the object recognition communities, such as the face verification [59] and person Re-ID [15], which is formulated as follows:

$$\mathcal{R} = \max\{\mathcal{M} + \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2, 0\}, \tag{3}$$

where $\mathcal{M}$ is the margin between positive pairs and negative pairs in the distance space, and $f(\cdot)$ is the learned feature mapping function. As a result, the relative distance between positive pairs and negative pairs are maximized, which is benefit to learn a deep ranking model in distinguishing the different individuals. As shown in Fig. 3, we argue that there are two drawbacks of directly applying this metric to solve the person Re-ID problem in deep learning framework, namely the equivalence of training samples and asymmetric gradient back-propagation, which will significantly weaken the generalization ability of the learned deep ranking model on the testing data.

Self-paced regularizer term In the SPL theory, a self-paced regularizer term is introduced to adaptively update the weights of samples according to both the training loss and model age. As shown in Fig. 3 (a), the easy samples will contribute more than the hard samples when the model is young, and all the samples will be involved equally when the model is mature. For this purpose, we propose a novel soft polynomial regularizer term, which is formulated as follows:

$$\mathcal{G} = \lambda(\frac{1}{t}\|\mathbf{u}\|_2^t - \frac{1}{\vartheta}\sum_{i=1}^{N} u_i), \tag{4}$$

where $\lambda > 0$ is the model age, $1 > \vartheta > 0$ is the mature age, and t is the polynomial order. Different from the recent self-paced regularizers, such as hard weighting [24] and soft weighting [56], our method penalizes the loss according to the value of polynomial order. As a result, the weighting scheme deduced by our regularizer term can approach all of them.

Symmetric regularizer term The goal of our symmetric regularizer term is to revise the asymmetric gradient back-propagation deduced by the relative similarity comparison metric. As a result, the intra-class distance can be minimized and the inter-class distance can be maximized simultaneously in each triplet unit, as shown in Fig. 3 (c). We penalize the deviation between two negative distances to keep the symmetric gradient back-propagation, which is formulated as follows:

$$\mathcal{S} = \frac{1}{\gamma}\log\left(1 + \exp\left(\gamma\mathcal{Z}\right)\right), \tag{5}$$

where $\mathcal{Z} = \left|\|f(\mathbf{x}_i^p) - f(\mathbf{x}_i^n)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2\right|$ is the deviation measured in the Euclid distance, and γ is the sharpness parameter. As shown in Fig. 3 (b), we introduce F_1 and F_3 to jointly revise the back-propagation of negative sample and positive sample in each triplet unit. What's more [1], the strength

[1]For the detail analysis of how to control the direction, please refer to Eq. (10) in the optimization section.

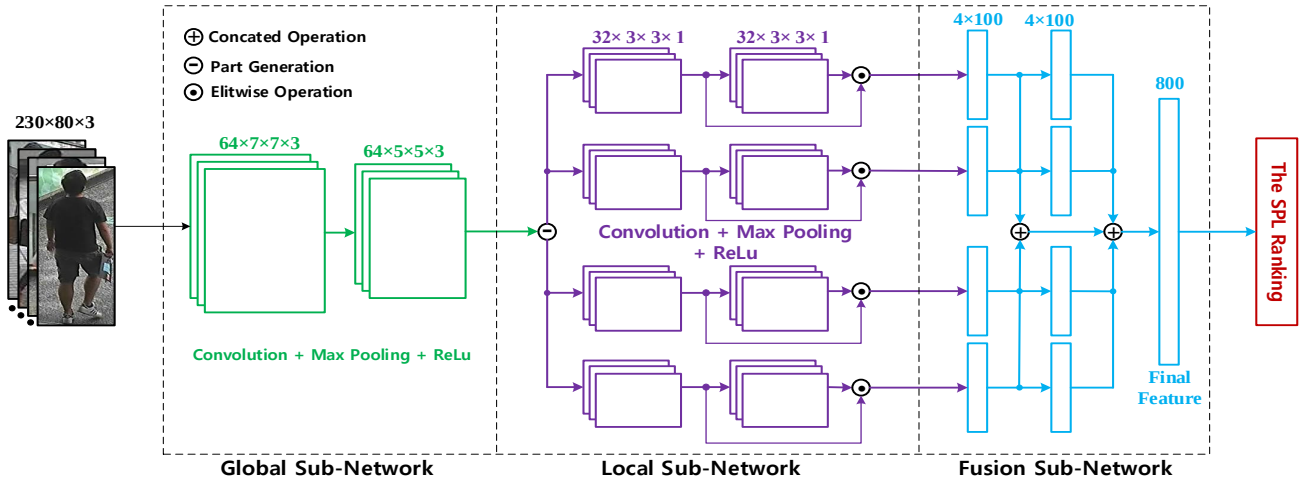

Figure 4: The deep feature learning and fusion network. This architecture is comprised of three sub-networks: glob sub-network, local sub-network and fusion sub-network. The first two parts extract the global feature representations and local feature representations from person images by using convolutional layers, max-pooling layers and part generation strategy. The third parts learns and fuses the local feature representations from the second part by using fully connected layers. Finally, the concated feature representations are fed into the loss layer for similarity comparison.

and direction can be adaptively tuned according to the deviation.

Parameter regularizer term In order to smooth the parameters of entire neural network, we define the following regularizer term:

$$\mathcal{P} = \sum_{k=1}^{K} \|\mathbf{W}^k\|_F^2 + \|\mathbf{b}^k\|_2^2 , \tag{6}$$

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

3.2. Deep Neural Network

In order to incorporate feature extraction and metric learning into an end-to-end framework, we propose a novel deep neural network which applies the part strategy to learn and fuse features from each individual. As shown in Fig. 4, the network is consisted of three subnetworks, which are introduced in the following paragraphs.

Global subnetwork It takes images in size of $230 \times 80 \times 3$ as input, and passes through two 64 learned filters of size $7 \times 7 \times 3$ and $5 \times 5 \times 3$, respectively. Then, the resulting feature maps are passed through a max pooling kernel of size 3×3 with stride 3. Finally, these feature maps are passed through a rectified linear unit (ReLU).

Local subnetwork We firstly divided 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. The outputs of the first and second local convolutional layer are summarized using eltwise operation. Afterwards, the resulting feature maps are passed through a max pooling kernel of size 3×3 with stride 1. Finally, we add a rectified linear unit (ReLU) after each max pooling layer.

Fusion subnetwork It takes local feature maps of different body parts as input, and learns discriminative features by concatenating two fully connected layers in each team. The dimension of the fully connected layers is 100 and a rectified linear unit (ReLU) is added between them. Then, the resulting features of the first four fully connected layers are concatenated to be fused by adding another fully connected layer in dimension of 400. Finally, the one 400 and four 100 dimensional features are concatenated to further generate the output 800 dimensional features.

3.3. Optimization

We use the gradient back-propagation method to optimize the parameters of deep CNN and weights of training samples.

For simplicity, we consider the deep parameters as a whole and define $\mathbf{\Omega}^k = [\mathbf{W}^k, \mathbf{b}^k]$ and $\mathbf{\Omega} = \{\mathbf{\Omega}^1, \cdots, \mathbf{\Omega}^K\}$.

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

$$\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}} = \sum_{i=1}^{N} u_i r(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) + \zeta s(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n) + 2\xi \sum_{k=1}^{K} \mathbf{\Omega}^k, \quad (7)$$

where the three terms represent gradient of the relative similarity term, the symmetric regularizer term and the parameter regularizer term, respectively.

We define $\mathcal{T} = \mathcal{M} + \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2$, therefore the gradient of relative similarity term can be formulated as follows:

$$r = \begin{cases} \frac{\partial \mathcal{R}(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)}{\partial \mathbf{\Omega}}, & if\ \mathcal{T} > 0; \\ 0, & else. \end{cases}, \quad (8)$$

where $\frac{\partial \mathcal{R}}{\partial \mathbf{\Omega}}$ is formulated as follows:

$$\begin{aligned} \frac{\partial \mathcal{R}}{\partial \mathbf{\Omega}} = & 2(f(\mathbf{x}_i^a) - f(\mathbf{x}_i^p))' \cdot \frac{\partial f(\mathbf{x}_i^a) - \partial f(\mathbf{x}_i^p)}{\partial \mathbf{\Omega}} \\ & -2(f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n))' \cdot \frac{\partial f(\mathbf{x}_i^a) - \partial f(\mathbf{x}_i^n)}{\partial \mathbf{\Omega}} \\ & -2(f(\mathbf{x}_i^p) - f(\mathbf{x}_i^n))' \cdot \frac{\partial f(\mathbf{x}_i^p) - \partial f(\mathbf{x}_i^n)}{\partial \mathbf{\Omega}}. \end{aligned} \quad (9)$$

By defining $\mathcal{D} = \|f(\mathbf{x}_i^p) - f(\mathbf{x}_i^n)\|_2^2 - \|f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n)\|_2^2$, then the gradient of symmetric regularizer term can be formulated as follows:

$$s = \eta \,\mathrm{sign}(\mathcal{D}) \cdot \frac{\partial \mathcal{D}(\mathbf{x}_i^a, \mathbf{x}_i^p, \mathbf{x}_i^n)}{\partial \mathbf{\Omega}}, \quad (10)$$

where $\eta = \exp(\gamma \mathcal{Z})/(1 + \exp(\gamma \mathcal{Z}))$ and $\mathrm{sign}(\mathcal{D})$ denote the strength and direction in the symmetric back-propagation, and $\frac{\partial \mathcal{D}}{\partial \mathbf{\Omega}}$ is formulated as follows:

$$\begin{aligned} \frac{\partial \mathcal{D}}{\partial \mathbf{\Omega}} = & 2(f(\mathbf{x}_i^a) - f(\mathbf{x}_i^n))' \cdot \frac{\partial f(\mathbf{x}_i^a) - \partial f(\mathbf{x}_i^n)}{\partial \mathbf{\Omega}} \\ & -2(f(\mathbf{x}_i^p) - f(\mathbf{x}_i^n))' \cdot \frac{\partial f(\mathbf{x}_i^p) - \partial f(\mathbf{x}_i^n)}{\partial \mathbf{\Omega}}. \end{aligned} \quad (11)$$

As shown in Fig. 5, the deduced strength and direction in controlling the gradient back-propagation can be adaptively updated according to the distance derivation, which is benefit to promote the symmetric back-propagation.

In order to update the weights of samples in each iteration, we deduce the closed form solution of the SPL model under the proposed regularizer term. Because the soft polynomial regularizer is convex in $[0, 1]$, it is easy to derive the optimal solution to $\min_{\mathbf{u} \in [0,1]} \sum_{i=1}^{N} u_i \mathcal{R} + \mathcal{G}(\mathbf{u}, \lambda, \vartheta)$ as follows:

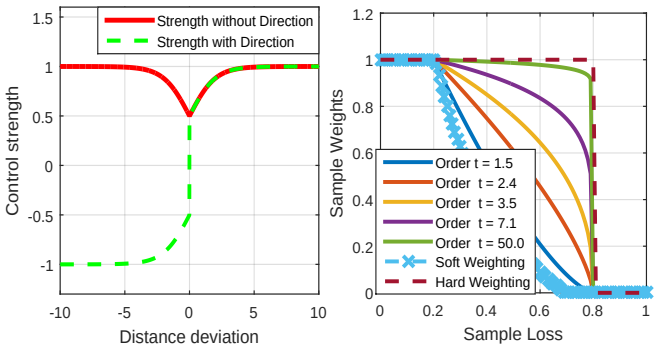

Figure 5: Illustration of the symmetric back-propagation and comparison of different weighting schemes, in which the left one shows that the deduced strength and direction in controlling the gradient back-propagation can be adaptively updated according to the distance derivation, and the right one shows that our method can approach nearly all the weighting schemes by just tuning the polynomial order.

$$u_i^* = \begin{cases} 1, & if\ \mathcal{R} < \lambda(\frac{1}{\vartheta} - 1), \\ 0, & if\ \mathcal{R} > \frac{\lambda}{\vartheta}, \\ (\frac{1}{\vartheta} - \frac{R}{\lambda})^{1/(t-1)}, & otherwise. \end{cases}. \quad (12)$$

The comparison with hard and soft weighting schemes are shown in Fig. 5, in which our method can approach them by tuning the polynomial order. If the loss is smaller than a threshold λ/ϑ, it will be treated as an easy sample and assigned a positive weight; If the loss is further smaller $\lambda(1/\vartheta - 1)$, the sample is treated as a faithful sample weighted by 1. Therefore, the easy-sample-first property [24] and soft weighting strategy [56] are all inherited in our method.

From the above derivations, we can see that the deep parameters and sample weights can be easily optimized given the values of $f(\mathbf{x}_i^a)$, $f(\mathbf{x}_i^p)$, $f(\mathbf{x}_i^n)$ and $\partial f(\mathbf{x}_i^a)/\partial \mathbf{\Omega}$, $\partial f(\mathbf{x}_i^p)/\partial \mathbf{\Omega}$, $\partial f(\mathbf{x}_i^n)/\partial \mathbf{\Omega}$, in which they can be obtained by separately running the forward and backward propagation by traveling all the triplets in each mini-batch. As the algorithm needs to accumulate the gradients in a self-paced way, we call it the self-paced gradient descent algorithm. We show the process in Algorithm 1.

Algorithm 1 The self-paced gradient descent algorithm

Input:

The training triplets $\mathbf{X}$, learning rate τ, age updating rate ω, maximum iterations H, margin parameters $\mathcal{M}$, age parameters λ, ϑ, weight parameters ζ, ξ and sharpness parameter γ.

Output:

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

repeat

1. Given an anchor sample $\mathbf{x}_i^a$, we randomly generate 200 triplets for each anchor in a mini-batch. Then, we calculate the output features of $f(\mathbf{x}_i^a)$, $f(\mathbf{x}_i^p)$ and $f(\mathbf{x}_i^n)$ of all the triplets by forward propagation.

repeat

a) Update the sample weights parameters $\mathbf{u}$ according to Eq. (3) and Eq. (12);

b) Calculate $\frac{\partial \mathcal{R}}{\partial \mathbf{\Omega}}, \frac{\partial \mathcal{D}}{\partial \mathbf{\Omega}}$ according to Eq. (9) and Eq. (11);

c) Increment the gradient $\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}}$ according to Eq. (7);

until Travel all the triplet units in each mini-batch.

2. Update the deep parameter $\mathbf{\Omega}_{h+1} = \mathbf{\Omega}_h - \tau_h \frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}_h}$ and model age $\lambda_{h+1} = \lambda_h / \omega_h$, with $h \leftarrow h + 1$.

until $h > H$

4. Experiment

In this section, we firstly introduce the datasets, the parameter setting and the evaluation protocol. Then, we evaluate the performance of our approach on five benchmark datasets, respectively. Finally, we give a detailed analysis of the experimental results.

4.1. Datasets and Settings

Datasets: Our experiments were conducted on five public datasets: the VIPeR [9], 3DPeS [60], CUHK01 [32], CUHK03 [45] and Market1501 [61]. Specially, VIPeR has 632 pedestrian image pairs captured outdoor with varying viewpoints and illumination conditions. 3DPeS contains 1011 image of 192 pedestrians captured from 8 outdoor camera views with significantly different viewpoints. CHUK01 contains 971 pedestrians from two disjoint camera views. CUHK03 has 13164 images of 1360 pedestrians captured by six different cameras. Market1501 contains 32668 images of 1501 identities, in which each identity is captured by six cameras at most and two cameras at least. Each pedestrian has at least one sample under each camera view.

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 $\tau = 0.01$, age updating rate $\omega = 0.9$, weight parameters $\zeta = 0.1, \xi = 0.01$, sharpness parameter $\gamma = 0.9$, age parameters $\lambda = 0.6, \vartheta = 0.75$ and margin parameter $\mathcal{M} = 1.1$.

Evaluation protocol: Our experiment follows the single-shot protocol in [15], in which 316 pedestrians in the VIPeR dataset, 96 pedestrians from the 3DPeS dataset, and 871 pedestrians of the CUHK01 dataset are randomly chosen to train the network, and the others are used to evaluate the performance. For the CUHK03 and Market1501 datasets, we use the provided fixed training and testing set in our experiment. The performance on the CUHK03 dataset is evaluated under the same single-shot protocol, and the performance on the Market1501 dataset is evaluated under both the single-query and multi-query evaluation settings as in [62]. The cumulative matching characteristic (CMC) curve is used to measure the performance of each method, which is an estimation of finding the corrected top n match. Besides, the mAP is also used to evaluate the performance on the Market1501 dataset. To obtain the statistical results, we repeated the testing 10 times and reported the average results.

4.2. Results

We compare our results with the following methods, namely the Quadruplet [63], Bow [61], kLFDA [8], SCSP [65], FPNN [45], LDNS [62], JSC [42], LMNN [66], LOMO+XQDA [25], IDLA [14], CDVM [67], LMF+LADF [10], KISSME [31], LSSCDL [68] and ME [64]. In order to show how much our DSPL method contributes to the performance,

Table 1: Matching rates(%) on the VIPeR dataset.

Methods	Top1	Top5	Top10	Top15	Top20
LOMO+XQDA [25]	40.00	68.13	80.51	87.37	91.08
Quadruplet [63]	49.05	73.10	81.96	−−	−−
ME [64]	45.89	77.40	88.87	93.52	95.84
LMF+LADF [10]	43.39	73.04	84.87	90.85	93.70
SCSP [65]	53.54	82.59	91.49	**95.09**	**96.65**
Our method (Baseline)	45.57	68.67	78.48	81.65	83.86
Our method (DSPL)	**56.32**	**83.04**	**92.01**	93.78	95.88

Table 2: Matching rates(%) on the 3DPeS dataset.

Methods	Top1	Top5	Top10	Top15	Top20
KISSME [31]	22.94	48.71	62.21	72.39	78.11
LF [35]	33.43	45.50	69.98	76.53	81.03
ME [64]	53.30	76.79	86.03	89.37	92.78
kLFDA [8]	54.02	77.74	85.92	90.04	92.38
SCSP [65]	57.29	78.97	85.01	89.52	91.51
Our Method (Baseline)	63.38	85.42	93.21	93.78	95.07
Our Method (DSPL)	**72.23**	**90.69**	**95.34**	**96.78**	**97.51**

we take the results of relative similarity comparison metric as baseline. Comparison results on the five datasets are shown in Table 1 to Table 5, respectively. In these tables, the best performance is highlighted in red, and the second best is is highlighted in blue.

For the VIPeR dataset, we compare our method with both the traditional methods and the deep learning based method, as shown in Table 1. From the results, we can see that our DSPL outperforms the baseline method with 10.75% in Top 1 accuracy, which demonstrates its effective by introducing the SPL and symmetric gradient back-propagation regularization. What's more, it outperforms the previous best performed method SCSP [65] with 2.78% in Top 1 accuracy.

Table 2 lists the results on the 3DPeS dataset, in which our baseline method gets the second best performance, contributed by the part-based deep CNN architecture, and our DSPL method achieves the best performance in all Top 1 to Top 20 accuracies. Compared with previous best performed method SCSP [65] on this dataset, our two methods outperform it by 6.09% and 14.94% in Top 1 accuracy, respectively. In addition, benefit from the SPL strategy and symmetric gradient back-propagation constraint used in our method, the DSPL method wins the baseline method 8.85% in Top 1 accuracy.

Table 3: Matching rates(%) on the CUHK01 dataset.

Methods	Top1	Top5	Top10	Top15	Top20
KISSME [31]	29.40	59.34	71.45	80.09	88.12
LMNN [66]	21.17	49.49	61.12	69.93	78.32
IDLA [14]	65.00	89.33	92.04	93.74	96.51
JSC [42]	65.71	89.41	92.52	93.74	96.63
CDVM [67]	66.50	93.00	96.50	99.00	99.00
Our method (Baseline)	71.14	89.12	94.19	97.85	98.84
Our method (DSPL)	**81.33**	**94.35**	**98.23**	**100.00**	**100.00**

In Table 3, we report the comparison results with the state-of-the-art methods on the CUHK01 dataset, in which our DSPL method achieves the best performance in all comparison groups from Top 1 to Top 20. Specially, our baseline method outperforms the previous best method CDVM [67] with 4.64% in Top 1, which demonstrates the effective by applying the part-based CNN to extract the feature representations. What's more, our DSPL method outperforms the baseline method and CDVM method with 10.19% and 14.83% in Top 1, respectively.

For the CUHK03 dataset, we compare our method with several state-of-the-art methods. The detail results are shown in Table 4, in which our DSPL method also achieves the best performance in all comparison groups from Top 1 to Top 20. Compared with the previous best method CDVM [67], our baseline method fall behind it by 1.08% in Top 1 accuracy, while our DSPL method outperforms it by 14.77% in Top 1 accuracy. Benefit from the DSPL and symmetric regularizer used in our method, the DSPL method wins the baseline method by 15.85% in Top 1 accuracy.

Finally, the Market1501 dataset is a newly proposed large scale dataset for person Re-ID. The best performance was obtained by a conventional method LDNS [62]. As illustrated in Table 5, the proposed two methods outperform LDNS by 2.96% and 11.87% in Top 1 accuracy under the single-query setting,

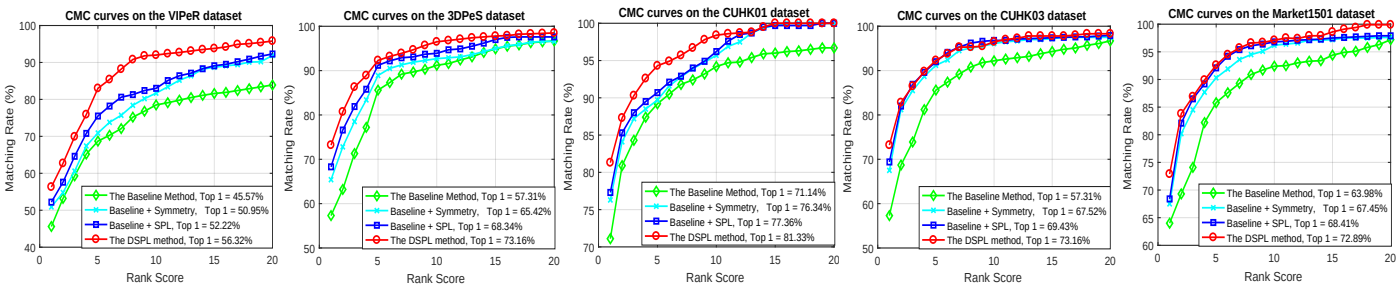

Figure 6: Comparison results of each component that contributes to the final performance. From these results, we can conclude that the SPL training strategy and symmetric constraint can continuously improve the identification performance and our DSPL method can get the best performance by incorporating the two constraints into an end-to-end learning framework.

Table 4: Matching rates(%) on the CUHK03 dataset.

Methods	Top1	Top5	Top10	Top15	Top20
LOMO+XQDA [25]	52.20	81.29	90.94	94.21	95.01
FPNN [45]	20.65	51.02	68.83	76.38	81.45
IDLA [14]	54.74	87.59	94.01	95.02	95.41
LSSCDL [68]	57.00	84.38	90.93	94.32	95.12
CDVM [67]	58.39	85.56	92.57	94.48	96.60
Our method (Baseline)	57.31	85.62	92.19	94.85	96.74
Our method (DSPL)	**73.16**	**92.26**	**96.54**	**97.75**	**98.45**

Table 5: Matching rates(%) on the Market1501 dataset.

Methods	Single-Query		Multi-Query	
	Top1	mAP	Top1	mAP
Bow [61]	34.38	14.10	42.64	19.47
kLFDA [8]	51.37	24.43	52.67	27.36
KISSME [31]	40.50	19.02	−−	−−
LDNS [62]	61.02	35.68	71.56	46.03
SCSP [65]	51.90	26.35	−−	−−
Our Method (Baseline)	**63.98**	**38.21**	**79.52**	**53.01**
Our Method (DSPL)	**72.89**	**46.68**	**87.05**	**57.41**

and 7.96% and 15.49% in Top 1 accuracy under the multi-query setting, respectively. Again, our DSPL method wins the baseline method by 8.91% and 7.53% in Top 1 accuracy under the single-query and the multi-query evaluation settings, respectively. For the mAP evaluation, the same conclusion can be made.

Table 6: Matching rates by using different networks.

Networks	VIPeR	3DPeS	CUHK01	CUHK03	Market501
G-Net	39.28	50.31	54.35	42.12	45.54
P-Net	**56.32**	**72.23**	**81.33**	**73.16**	**72.89**

4.3. Analysis

In this section, we analyze the experimental results of our method from two aspects, namely the effectiveness of each contribution to the final performance and the robustness of our method to different parameter settings. The detailed analysis results are given in the following paragraphs.

Effectiveness of each contribution Firstly, we report some intermediate results of each contribution in our method, namely the baseline method, the baseline method + symmetric regularizer, the baseline method + SPL and the DSPL method, so as to illustrate how much each of them contributes to the ranking performance on the five datasets, respectively. The comparison results are shown in Fig. 6, from which we can make the following two conclusions: 1) The baseline method can be improved by separately incorporating the SPL strategy or the symmetric regularizer into the relative similarity comparison framework; 2) The DSPL method significantly outperforms the baseline method and the other two methods by jointly introducing the SPL strategy and symmetric regularizer into an end-to-end learning framework. The above two points tell us that it is very important to distinguish the reliability of samples and revise the gradient back-propagation in the training process.

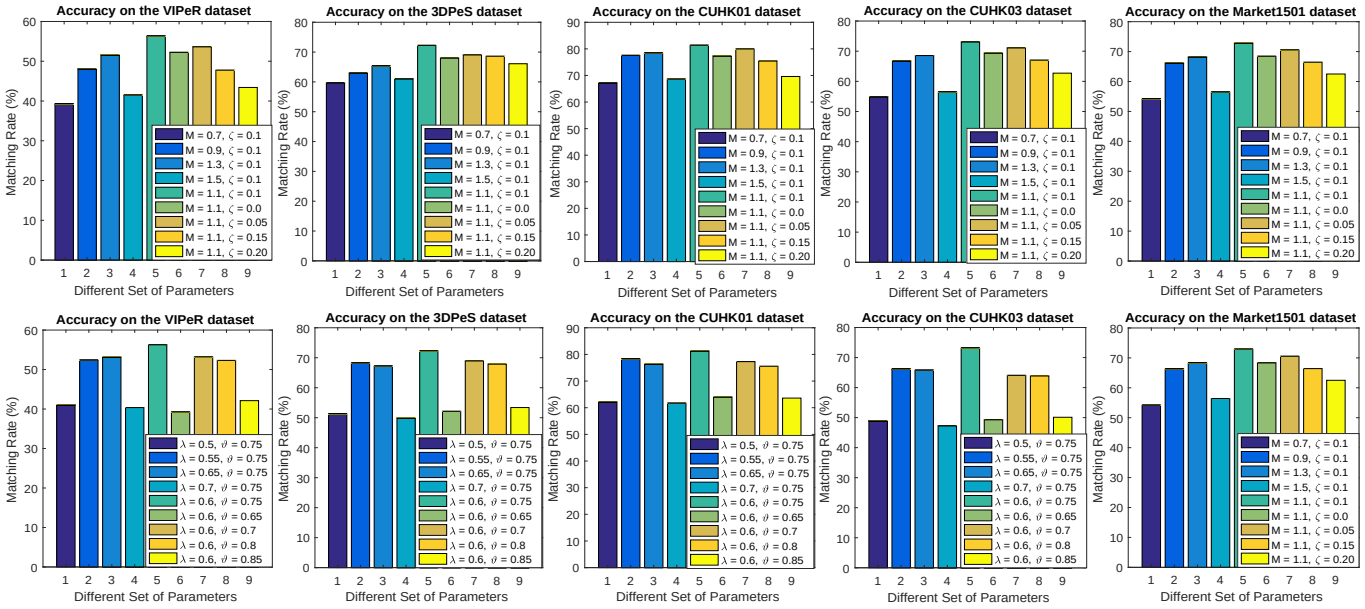

Figure 7: Influences of different parameter settings to the final performance. Specially, the first column shows how the symmetric regularizer term effects the performance, in which our DSPL method get its best performance by setting $\mathcal{M} = 1.1$, $\zeta = 0.1$. The second column shows how the SPL strategy effects the performance, in which our DSPL method get the promising results by setting $\lambda = 0.6$ and $\vartheta = 0.75$ on the five benchmark datasets, respectively.

Table 7: Matching rates of the SPL with different loss functions.

Metrics	VIPeR	3DPeS	CUHK01	CUHK03	Market501
CL	42.35	54.81	67.61	51.43	57.61
CL+SPL	**50.81**	**62.49**	**74.05**	**65.38**	**66.28**
TL	45.57	57.31	71.14	57.31	63.98
TL+SPL	**55.22**	**64.04**	**77.38**	**69.43**	**68.41**
OurLoss	50.95	65.42	73.64	67.52	67.45
OurLoss+SPL	**56.32**	**73.16**	**81.33**	**73.16**	**72.89**

Secondly, we report two set of comparison results to evaluate the superiority of our part-based neural network and the effectiveness of the SPL strategy with different distance metrics. For fair comparison, we keep other parts the same when compare one specified part in all these experiments. In order to evaluate the superiority of our part-based network, we build another global network, in which we get rid of eight small convolutional layers in the local subnetwork and take two large convolutional layers to replace them. The comparison results are shown in Table 6, in which our part-based neural network outperforms the global neural network in Top 1 accuracy on all the five datasets. The reason may come from two aspect: 1) Different body parts have different importance in representing the person appearance [14], and the part-based neural network allows to learn different body parts discriminatively; 2) Dividing the feature maps into different parts is a kind of data augmentation, and the data augmentation is a common way to improve the network performance in deep learning community. Besides, we evaluate the effectiveness of the SPL strategy with three different distance metrics, namely the contrastive loss, the triplet loss and the proposed loss. The comparison results are shown in Table 7, in which the SPL strategy improves the baseline performance of different metrics on all the five datasets. According to our experience, the reason of why the SPL works is due to the data distribution. As a fine-grained recognition task, person images usually gather together in feature space when the representation ability of the deep network is weak. At this time, easy samples are more beneficial to steadily enhance the representation ability of neural network. When the representation ability of deep model reaches a certain level, all the samples will be involved into the training process to boost the final performance.

Parameter influence To the best of our knowledge, the margin parameter $\mathcal{M}$, the weight parameter ζ and the age parameters λ, ϑ have major effects to the final ranking perfor-

mance in our DSPL method. The margin parameter $\mathcal{M}$ and weight parameter ζ jointly control the symmetric gradient back-propagation of the relative distance metric, and the age parameters λ, ϑ control the way of hard samples are involved into the training process. In the following, we give an empirical analysis of our method on the five datasets, respectively.

The results are shown in Fig. 7, in which our method achieves its best performance by setting $\mathcal{M} = 1.1, \zeta = 0.1$ and $\lambda = 0.6, \vartheta = 0.75$. We demonstrate the results from the following four points: 1) Small margin $\mathcal{M}$ will make the candidate positive and negative samples indistinguishable in the distance space; while a large margin will lead to the numerical instability problem. 2) Small weight ζ will weaken the symmetric constraint, which can lead to the asymmetric gradient back-propagation; while large weight will enhance the symmetric constraint, which is also harmful to computational stability. 3) The meaning of λ is the current age of model, and small λ will hinder the hard samples involved into the training process, while large λ will lead the easy samples and hard samples indistinguishable in the training process. 4) The meaning of ϑ is the mature age of model, and small ϑ will lead the hard samples involve into the training process too early, while large ϑ will make only a small amount of hard samples are involved into the training process. Therefore, we choose a moderate $\mathcal{M}$, ζ, λ and ϑ for the high-quality performance of the symmetric gradient back-propagation and self-paced training process.

Some ranking examples To obtain more insight of the D-SPL method, some ranking examples on the five benchmark datasets are shown in Fig. 8, in which person in green rectangle denotes the query image and person in red rectangle represents the matched candidate. For each dataset, we give the comparison ranking results of two methods, namely the baseline method and the DSPL method, in the first and the second column respectively. From these examples, we can conclude our DSPL performs much better the baseline method. The main reason is that our DSPL method applies the SPL strategy and symmetric regularization in the training process, which have been effective to improve the person Re-ID performance. Compared with

Figure 8: The ranking results on the five benchmark datasets in single-shot evaluation, in which person in green rectangle denotes the query image and person in red rectangle represents the matched candidate.

the ranking results on the CUHK01 and CUHK03 dataset, we find that our method can not always find the correct matches. In the future, we will strive to find an optimal saliency detection modular in our neural network, so as to further improve the ranking performance of cases as shown the CUHK01 and CUHK03 datasets.

5. Conclusion

In this paper, we propose a novel person re-identification method by incorporating the SPL strategy and symmetric regularizer to perform integrated feature learning and fusion in an end-to-end deep framework. In order to extract the stable and discriminative features, we build a part-based neural network, in which the features of different body parts are first discriminately learned in the lower convolutional layers and then fused in the higher fully connected layers. The output features are further fed into the relative similarity comparison metric to optimize the deep parameters in gradient back-propagation. By introducing the SPL strategy into the distance metric, the sides effects of noisy samples or outliers can be alleviated by using a soft polynomial regularizer to adaptively up-

date the sample weights in each iteration. The asymmetric gradient back-propagation is revised by introducing the symmetric regularizer, therefore the intra-class distance is minimized and inter-class distance is maximized in each triplet unit. Extensive experimental results on the VIPeR, 3DPeS, CUHK01, CUHK03 and Market1501 datasets have shown that our method outperforms most of the state-of-the-art approaches in person re-identification.

References

[1] S. Zhou, J. Wang, J. Wang, Y. Gong, N. Zheng, Point to set similarity based deep feature learning for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[2] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[3] C. Su, S. Zhang, F. Yang, G. Zhang, Q. Tian, W. Gao, L. S. Davis, Attributes driven tracklet-to-tracklet person re-identification using latent prototypes space mapping, Pattern Recognition 66 (2017) 4–15.

[4] X. Liu, H. Wang, J. Wang, X. Ma, Person re-identification by multiple instance metric learning with impostor rejection, Pattern Recognition 67 (2017) 287–298.

[5] T. Li, H. Chang, M. Wang, B. Ni, R. Hong, S. Yan, Crowded scene analysis: A survey, IEEE transactions on circuits and systems for video technology 25 (3) (2015) 367–386.

[6] L. Ren, J. Lu, J. Feng, J. Zhou, Multi-modal uniform deep learning for rgb-d person re-identification, Pattern Recognition 72 (2017) 446–457.

[7] X. Ma, X. Zhu, S. Gong, X. Xie, J. Hu, K.-M. Lam, Y. Zhong, Person re-identification by unsupervised video matching, Pattern Recognition 65 (2017) 197–210.

[8] 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.

[9] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, Computer Vision–ECCV 2008 (2008) 262–275.

[10] 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.

[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. Saul, Distance metric learning for large margin nearest neighbor classification, Advances in neural information processing systems 18 (2006) 1473.

[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: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.

[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] 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.

[17] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person re-identification, arXiv preprint arXiv:1611.05666.

[18] S. Zhou, J. Wang, R. Shi, Q. Hou, Y. Gong, N. Zheng, Large margin learning in set to set similarity comparison for person re-identification, arXiv preprint arXiv:1708.05512.

[19] L. Wu, C. Shen, A. van den Hengel, Deep linear discriminant analysis on fisher networks: A hybrid architecture for person re-identification, Pattern Recognition 65 (2017) 238–250.

[20] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, Y. Wu, Learning fine-grained image similarity with deep ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1386–1393.

[21] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proceedings of the 22nd ACM international conference on Multimedia, ACM, 2014, pp. 675–678.

[23] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, arXiv preprint arXiv:1603.04467.

[24] M. P. Kumar, B. Packer, D. Koller, Self-paced learning for latent variable models, in: Advances in Neural Information Processing Systems, 2010, pp. 1189–1197.

[25] S. Liao, Y. Hu, X. Zhu, S. Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.

[26] 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.

[27] 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.

[28] R. Zhao, W. Oyang, X. Wang, Person re-identification by saliency learning, IEEE transactions on pattern analysis and machine intelligence 39 (2) (2017) 356–370.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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

[37] 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.

[38] D. Chen, Z. Yuan, J. Wang, B. Chen, G. Hua, N. Zheng, Exemplar-guided similarity learning on polynomial kernel feature map for person re-identification, International Journal of Computer Vision (2017) 1–23.

[39] 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.

[40] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, Q. Mary, Person re-identification by support vector ranking., in: BMVC, Vol. 2, 2010, p. 6.

[41] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.

[42] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1288–1296.

[43] R. Zhang, L. Lin, R. Zhang, W. Zuo, L. Zhang, Bit-scalable deep hashing with regularized similarity learning for image retrieval and person re-identification, IEEE Transactions on Image Processing 24 (12) (2015) 4766–4779.

[44] 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.

[45] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.

[46] Y.-C. Chen, X. Zhu, W.-S. Zheng, J.-H. Lai, Person re-identification by camera correlation aware feature augmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence.

[47] Y. Yan, B. Ni, Z. Song, C. Ma, Y. Yan, X. Yang, Person re-identification via recurrent feature aggregation, in: European Conference on Computer Vision, Springer, 2016, pp. 701–716.

[48] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[49] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 41–48.

[50] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, A. Hauptmann, Self-paced learning with diversity, in: Advances in Neural Information Processing Systems, 2014, pp. 2078–2086.

[51] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE transactions on pattern analysis and machine intelligence 39 (5) (2017) 865–878.

[52] Y. J. Lee, K. Grauman, Learning the easy things first: Self-paced visual category discovery, in: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1721–1728.

[53] J. S. Supancic, D. Ramanan, Self-paced learning for long-term tracking, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 2379–2386.

[54] K. Tang, V. Ramanathan, L. Fei-Fei, D. Koller, Shifting weights: Adapting object detectors from image to video, in: Advances in Neural Information Processing Systems, 2012, pp. 638–646.

[55] H. Li, M. Gong, D. Meng, Q. Miao, Multi-objective self-paced learning, in: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI Press, 2016, pp. 1802–1808.

[56] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, A. G. Hauptmann, Self-paced learning for matrix factorization., in: AAAI, 2015, pp. 3196–3202.

[57] J. Liang, Z. Li, D. Cao, R. He, J. Wang, Self-paced cross-modal subspace matching, in: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval, ACM, 2016, pp. 569–578.

[58] L. Lin, K. Wang, D. Meng, W. Zuo, L. Zhang, Active self-paced learning for cost-effective and progressive face identification, IEEE transactions on pattern analysis and machine intelligence.

[59] F. Schroff, D. Kalenichenko, J. Philbin, Facenet: A unified embedding for face recognition and clustering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 815–823.

[60] D. Baltieri, R. Vezzani, R. Cucchiara, Sarc3d: a new 3d body model for people tracking and re-identification, Image Analysis and Processing–ICIAP 2011 (2011) 197–206.

[61] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: The IEEE International Conference on Computer Vision (ICCV), 2015.

[62] 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.

[63] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: A deep quadruplet network for person re-identification, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[64] S. Paisitkriangkrai, C. Shen, A. van den Hengel, Learning to rank in person re-identification with metric ensembles, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1846–1855.

[65] 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.

[66] K. Q. Weinberger, L. K. Saul, Distance metric learning for large margin nearest neighbor classification, Journal of Machine Learning Research 10 (Feb) (2009) 207–244.

[67] L. Lin, G. Wang, W. Zuo, X. Feng, L. Zhang, Cross-domain visual matching via generalized similarity measure and feature learning, IEEE transactions on pattern analysis and machine intelligence 39 (6) (2017) 1089–1102.

[68] Y. Zhang, B. Li, H. Lu, A. Irie, X. Ruan, Sample-specific svm learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1278–1287.