Point to Set Similarity Based Deep Feature Learning for Person Re-identification

Sanping Zhou, Jinjun Wang, Jiayun Wang, Yihong Gong, Nanning Zheng
The Institute of Artificial Intelligence and Robotic, Xi'an Jiaotong University

Abstract

Person re-identification (Re-ID) remains a challenging problem due to significant appearance changes caused by variations in view angle, background clutter, illumination condition and mutual occlusion. To address these issues, conventional methods usually focus on proposing robust feature representation or learning metric transformation based on pairwise similarity, using Fisher-type criterion. The recent development in deep learning based approaches address the two processes in a joint fashion and have achieved promising progress. One of the key issues for deep learning based person Re-ID is the selection of proper similarity comparison criteria, and the performance of learned features using existing criterion based on pairwise similarity is still limited, because only Point to Point (P2P) distances are mostly considered. In this paper, we present a novel person Re-ID method based on Point to Set similarity comparison. The Point to Set (P2S) metric can jointly minimize the intra-class distance and maximize the inter-class distance, while back-propagating the gradient to optimize parameters of the deep model. By utilizing our proposed P2S metric, the learned deep model can effectively distinguish different persons by learning discriminative and stable feature representations. Comprehensive experimental evaluations on 3DPeS, CUHK01, PRID2011 and Market1501 datasets demonstrate the advantages of our method over the state-of-the-art approaches.

1. Introduction

Given one single shot or multiple shots of a pedestrian from one camera view, person re-identification (Re-ID) aims to match the same person amongst a set of gallery candidates captured from the disjoint camera networks. It is an important task to many surveillance applications such as person association [25], multi-target tracking [39] and behavior analysis [14]. The problem is also very challenging, because the typical setup of video surveillance system in unconstrained environments usually generates significant appearance changes due to the variations in view angel, background clutter, illumination condition and mutual occlusion, as shown in Fig. 1. Therefore, a discriminative and stable feature representation should be learned to distinguish different individuals for person Re-ID, in which the intra-class distance is smaller than the inter-class ones.

Figure 1. The challenges to person re-identification problem in public space, where the query image in the probe set is denoted in green box and the matched images in the gallery set are shown in black box.

To address these challenges, extensive works have been reported in the past few years, which could be roughly divided into two categories: 1) developing robust descriptor to handle the variations in persons' appearance, and 2) designing effective distance metric to measure the similarity between persons' images. For the first category, different cues are employed for discriminative and stable feature extraction, and representative descriptors include the Local Binary Pattern (LBP) [35], Ensemble of Local Feature (ELF) [10] and Local Maximal Occurrence (LOMO) [41]. For the second category, labeled images are used to learn a distinctive distance metric, and popular metric learning methods include the Locally Adaptive Decision Function (LADF) [19], Large Margin Nearest Neighbor (LMNN) [32], Information Theoretic Metric Learning (ITML) [6], etc. Since both line of works regard feature extraction and metric learning processes as two disjoint steps, their performances are limited.

Recently, the deep learning based methods have been

proved to be effective for person Re-ID [1, 7, 23], because they can incorporate feature extraction and metric learning into an integrated framework, in which the two processes are implemented as two connected components: 1) a Deep Neural Network (DNN) to extract features from pedestrian images, and 2) a distance metric to compute the loss and back-propagate the gradients. Benefit from the powerful representation capability of the DNN, these methods have achieved the state-of-the-art performance on the benchmark datasets for person Re-ID [31, 34].

Despite the great success of these deep learning based methods achieved in person Re-ID, insufficient labeled training data has limited their generalization ability of learned models for the testing data, while collecting the training samples is quite labor intensive. Even though the triplet loss function [7] could effectively alleviate this problem by sampling a large set of anchor-positive-negative triplets, it is still based on the P2P distance such that the ranking performance of learned feature is still limited. In this paper, we propose a novel Point to Set distance metric to supervise a designed deep Convolutional Neural Network (CNN) to learn discriminative and stable feature representations for person Re-ID. In order to learn the feature representations from multiple perspectives, we construct an effective part-based deep CNN to extract discriminative features from different body parts of each person. The proposed framework is generic where different deep models, such as the AlexNet [16], VGGNet [30] or ResNet [11], could also be applied to extract feature representations from the input images. As a general loss, our proposed P2S metric can jointly minimize the intra-class distance and maximize the inter-class distance, while back-propagating the gradients to optimize parameters of the deep model. As demonstrated in our experiments, a large margin is held between the intra-class distance and inter-class distance in the learned feature space, such that its performance in differentiating the intra/inter-class person is superior than many state-of-the-art methods.

The main contributions of this work can be highlighted as follows: 1) A novel P2S distance metric is proposed to supervise a deep model to learn discriminative and stable feature representations for similarity comparison, which can penalize a large margin between the positive pairs and negative pairs in the learned feature space. Compared with the existing P2P distance based metrics, our method considers the P2S information and is more effective at improving the ranking performance; 2) An effective part-based deep CNN is constructed to extract discriminative and stable feature representations of different body parts for person Re-ID. The deep architecture is constituted of a global sub-network, a local sub-network and a fusion sub-network, such that different body parts are first discriminately learned in the global sub-network and local sub-network, and then fused in the fusion sub-network. Extensive experiments are conducted on several public benchmark datasets including 3DPeS, CUHK01, PRID2011 and Market1501, which show clear improvement of our method as compared with the state-of-the-art approaches.

2. Related Work

Extensive works have been reported to tackle the person Re-ID problem. These methods mainly focus on several different aspects of the issue such as developing robust feature descriptors, designing discriminative metrics and learning deep features. Blew we give a brief review of some representative ones.

Feature Designing Method The feature designing methods mainly focus on developing discriminative person representation which is robust to the cross view appearance variations. For examples, Zhao et al. [41] learned a mid-level filter from patch cluster to achieve cross view invariance. In [20], Liao et al. constructed a feature descriptor which analyzed the horizontal occurrence of local features and maximized the occurrence to make a stable representation against viewpoint changes. Ma et al. [22] presented the person image via covariance descriptor which was robust to illumination changes and background variations. In [8], Farenzena et al. augmented maximally stable color regions with histograms for person representation. Zhao et al. [40] learned the distinct salience feature to distinguish the matched person from others. In [5], Chen et al. employed a pre-learned pictorial structure model to localize the body parts more accurately. Wu et al. [33] 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 [15], Kviatkovsky et al. investigated the intra-distribution structure of color descriptor, which was invariant under certain illumination changes. Li et al. [17] matched person images observed in different camera views with complex cross-view transforms and applied it to person Re-ID.

Metric Learning Method The metric learning 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 instance, Zheng et al. [43] proposed a relative distance learning method from a probabilistic prospective. In [24], Mignon et al. learned a distance metric from sparse pairwise similarity constraints. Pedagadi et al. [28] utilized LADF to map the high dimensional features into a more discriminative low dimensional space. In [35], Xiong et al. further extended the LADF and several other metric learning methods by using kernel tricks and different regularizers. Nguyen et al. [26] measured the similarity of face pairs through cosine similarity, which was closely related to the inner product similarity. In [21], Loy et al. casted the

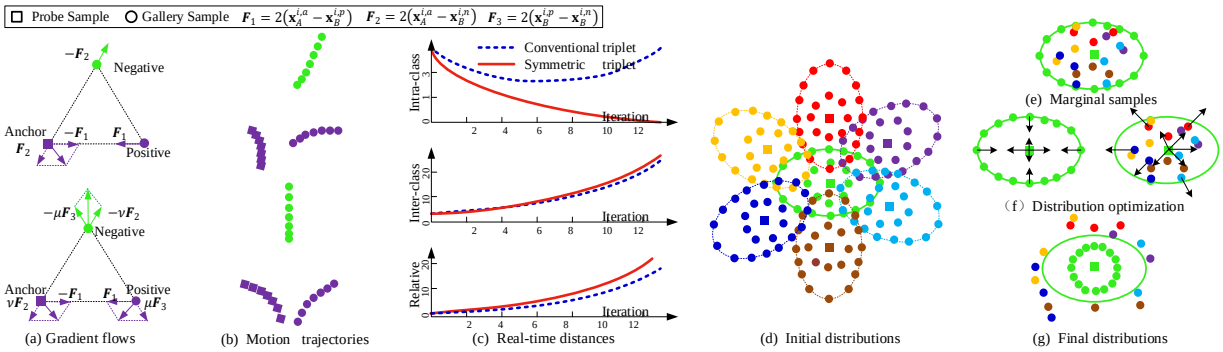

Figure 2. Illustration of the proposed P2S method, in which (a) shows the gradient flows of the conventional triplet formulation [7] and the proposed symmetric triplet formulation; (b) shows the two corresponding motion trajectories driven by the two gradient flows; (c) illustrates the changes of intra-class distances, inter-class distances and relative distances of the two triplet formulations with respect to iteration; (d) illustrates the initial distributions of samples in a mini-batch; (e) plots only the adaptively selected positive and negative samples; (f) shows the flow of distributions based on the proposed symmetric triplet formulation; and (g) shows the final distributions.

person Re-ID problem as an image retrieval task by considering the listwise similarity. Chen et al. [4] proposed a kernel based metric learning method to explore the nonlinearity relationship of samples in the feature space. In [13], Hirzer et al. learned a discriminative metric by using relaxed pairwise constraints. Prosser et al. developed [29] a ranking model using support vector machine.

Deep Learning Method As explained before, the deep learning based methods aim to incorporate the two into an integrated framework, in which adaptive feature representation can be learned under the supervision of distance metric. For example, Li et al. [18] 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 across views. In [1], Ahmed et al. proposed an improved deep learning framework which took pairwise images as inputs, and outputs a similarity value indicating whether the two input images depict the same person or not. Xiao et al. [34] proposed a domain guided dropout algorithm to improve the performance of deep CNN to extract robust feature representation for person Re-ID. In [36], Yi et al. constructed a siamese neural network to learn pairwise similarity, and used body parts to train the model. Ding et al. [7] applied the triplet loss to train the triplet deep framework for person Re-ID. In [31], Wang et al. proposed a unified triplet and siamese deep architecture which can jointly extract single-image and cross-image feature representations. Zhou et al. [44] propose an adaptive margin method to learn the deep features for person Re-ID in a siamese framework.

3. The Point to Set Model

Let $\mathbf{X} = \{\mathbf{X}_i\}_{i=1}^{N}$ be the input set of training samples, where $\mathbf{X}_i = \{\mathbf{x}_A^{i,a}, \mathbf{x}_B^{i,j}\}_{j=1}^{M}$ denotes the pairwise set of the i^{th} raw input data, N is the number of training identities, and M is the number of training images belonging to the i^{th} identity. The goal of our deep architecture is to learn 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:

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

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 similarity, we simplify the parameters of the neural network as a whole and define $\mathbf{W} = \{\mathbf{W}^1, \cdots, \mathbf{W}^K\}$ and $\mathbf{b} = \{\mathbf{b}^1, \cdots, \mathbf{b}^K\}$. The next sections depict our proposed metric.

3.1. The Point to Set Metric

The P2S metric is consisted of three terms, namely the pairwise term, the triplet term and the regularizer term, which can be formulated as follows:

$$\mathrm{L} = \mathrm{L_P}(\mathbf{X}, \mathbf{W}, \mathbf{b}) + \alpha \mathrm{L_T}(\mathbf{X}, \mathbf{W}, \mathbf{b}) + \beta \mathrm{R}(\mathbf{W}, \mathbf{b})\,, \tag{2}$$

where $\mathrm{L_P}$ is the pairwise term, $\mathrm{L_T}$ denotes the triplet term, R represents the regularizer term, and α, β are two constant weight parameters. Given an anchor sample, the pairwise term randomly selects positive and negative candidates to alleviate the overfitting problem, while the triplet term adaptively chooses the marginal samples to boost the ranking performance, and the regularizer term smoothes the parameters to preserve the numerical stability. These terms are elaborated in the next paragraphs.

The pairwise term To alleviate the overfitting problem, the pairwise term randomizes the selection of positive pairs and negative pairs to train the deep model. Specially, the

pairwise term aims to penalize when the positive distances are greater than a preset down-margin and the negative distances are smaller than a preset upper-margin. The hinge loss of the pairwise term can be formulated as follows:

$$\mathrm{L_P} = \frac{1}{\mathcal{Z}_p} \sum_{i,j=1}^{N} \sum_{r=1}^{M} \max\{\mathcal{C}_p - \mathrm{G}_{i,j}^{a}(\mathcal{M}_p - \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}\|_2^2), 0\}\,, \tag{3}$$

where $\mathcal{Z}_p$ is the normalization factor, the two parameters $\mathcal{M}_p > \mathcal{C}_p$ are used to define the down-margin and upper-margin, respectively. Specifically, $\mathcal{M}_p - \mathcal{C}_p$ represents the down-margin, and $\mathcal{M}_p + \mathcal{C}_p$ denotes the up-margin. Given the i^{th} and j^{th} identities, the indicator matrix $\mathrm{G}_{i,j}^{a}$ refers to the correspondence of the r^{th} image in camera B to the anchor image in camera A, which is defined as follows:

$$\mathrm{G}_{i,j}^{a} = \begin{cases} +1,\ if\ i = j,\ and\ r \leq M\,, \\ -1,\ if\ i \neq j,\ and\ r \leq M\,, \end{cases} \tag{4}$$

where $\mathrm{G}_{i,j}^{a}$ is in size of $N \times M$, and $\mathrm{G}_{i,j}^{a}(j,r) = 1$ means that the r^{th} image of the j^{th} identity is referred to the same person to that of the anchor image of the i^{th} identity, while $\mathrm{G}_{i,j}^{a}(j,r) = -1$ means the opposite.

Definition-*S*ymmetric Triplet:[1] Given a set of triplet training samples $\{\mathbf{x}_A^{i,a}, \mathbf{x}_B^{i,p}, \mathbf{x}_B^{i,n}\}_{i=1}^{N}$, in which $\{\mathbf{x}_A^{i,a}, \mathbf{x}_B^{i,p}\}$ is a positive pair and $\{\mathbf{x}_A^{i,a}, \mathbf{x}_B^{i,n}\}$ denotes a negative pair, the conventional triplet formulation penalizes a large relative margin $\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,n}\|_2^2 - \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,p}\|_2^2 \geq \mathcal{M}$ by using the loss $\mathrm{L} = \sum_{i=1}^{N} \max\{\mathcal{M} + \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,p}\|_2^2 - \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,n}\|_2^2, 0\}$. In our symmetric triplet formulation, we satisfy the above constraint by using the loss $\mathrm{L} = \sum_{i=1}^{N} \max\{\mathcal{M} + \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,p}\|_2^2 - [\mu\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{i,n}\|_2^2 + \nu\|\mathbf{x}_B^{i,p} - \mathbf{x}_B^{i,n}\|_2^2], 0\}$, where the first term denotes the intra-class distance, the second term and the third term are weighted to represent the inter-class distance, and μ, ν are two adaptive weights.

The triplet term The triplet term aims to improve the ranking performance by maximizing the relative distance between anchor to positive set and anchor to negative set. As illustrated in Fig. 2, we formulate the point to set distance as the average distance between anchor and marginal set samples, in which the anchor to negative set distance should also satisfy $\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}\|_2^2 < \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{k,s}\|_2^2$, where $i = j, i \neq k$ and $r, s \leq M$. Therefore by formulating the relative point to set distance in the novel symmetric triplet formulation, the hinge loss of the triplet term can be defined as follows:

$$\mathrm{L_T} = \frac{1}{\mathcal{Z}_t} \sum_{i,j,k=1}^{N} \sum_{r,s=1}^{M} \max\{\mathcal{M}_t - \mathrm{T}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s}), 0\}\,, \tag{5}$$

where $\mathcal{Z}_t$ is the normalization factor, $\mathcal{M}_t$ denotes the relative margin parameter, and $\mathrm{T}(\cdot)$ represents the relative point to set distance:

$$\mathrm{T} = \mathrm{P}_{i,j}^{a}\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}\|_2^2 - \mathrm{N}_{i,k}^{a}[\mu\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{k,s}\|_2^2 + \nu\|\mathbf{x}_B^{j,r} - \mathbf{x}_B^{k,s}\|_2^2]\,, \tag{6}$$

where $\mathrm{P}_{i,j}^{a}, \mathrm{N}_{i,j}^{a}$ denote the positive and negative indicator matrixes, and μ, ν are two adaptive weight parameters. Given the triplet identity $\{i, j, k\}$, the indicator matrixes $\mathrm{P}_{i,j}^{a}$ and $\mathrm{N}_{i,k}^{a}$ represent the matched and unmatched candidates of the r^{th} and s^{th} image in camera B to the anchor image in camera A, respectively. They are defined as follows:

$$\mathrm{P}_{i,j}^{a} = \begin{cases} 1,\ if\ i = j,\ and\ r = \tau_p(a)\,, \\ 0,\ else\,, \end{cases} \tag{7}$$

$$\mathrm{N}_{i,k}^{a} = \begin{cases} 1,\ if\ i \neq k,\ and\ s = \tau_n(a)\,, \\ 0,\ else\,, \end{cases} \tag{8}$$

where $\mathrm{P}_{i,j}^{a}$ and $\mathrm{N}_{i,k}^{a}$ are both in size of $N \times M$, and $\mathrm{P}_{i,j}^{a}(j,r) = 1$ means that the r^{th} image of the j^{th} identity is referred to the same person to that of the anchor image of the i^{th} identity, $\mathrm{N}_{i,k}^{a}(j,s) = 1$ means that the s^{th} image of the k^{th} identity is referred to the different person to that of the anchor image of the i^{th} identity, while $\mathrm{P}_{i,j}^{a}(j,r) = 0$ and $\mathrm{N}_{i,k}^{a}(j,s) = 0$ mean the opposite. The positive and negative marginal samples are represented by $\tau_p(a)$ and $\tau_n(a)$, and both of them can be collected by using the nearest neighbor search algorithm.

The regularizer term To smooth the parameters of the entire neural network, we define the following regularizer term:

$$\mathrm{R} = \sum_{k=1}^{K} \|\mathbf{W}^k\|_F^2 + \|\mathbf{b}^k\|_2^2\,, \tag{9}$$

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

3.2. Deep Architecture

The proposed P2S metric is combined with our proposed part-based deep CNN to implement an end-to-end framework for both feature learning and fusion. As shown in Fig. 3, the proposed deep architecture is consisted of three sub-networks: global sub-network, local sub-network and fusion sub-network. The following paragraphs explain the respective networks in more detail.

Global sub-network The first part of our network is a global sub-network, which is consisted of 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

[1] As shown in Fig. 2, the symmetric triplet formulation outperforms the conventional one by optimizing the gradient directions for the positive and the negative samples.

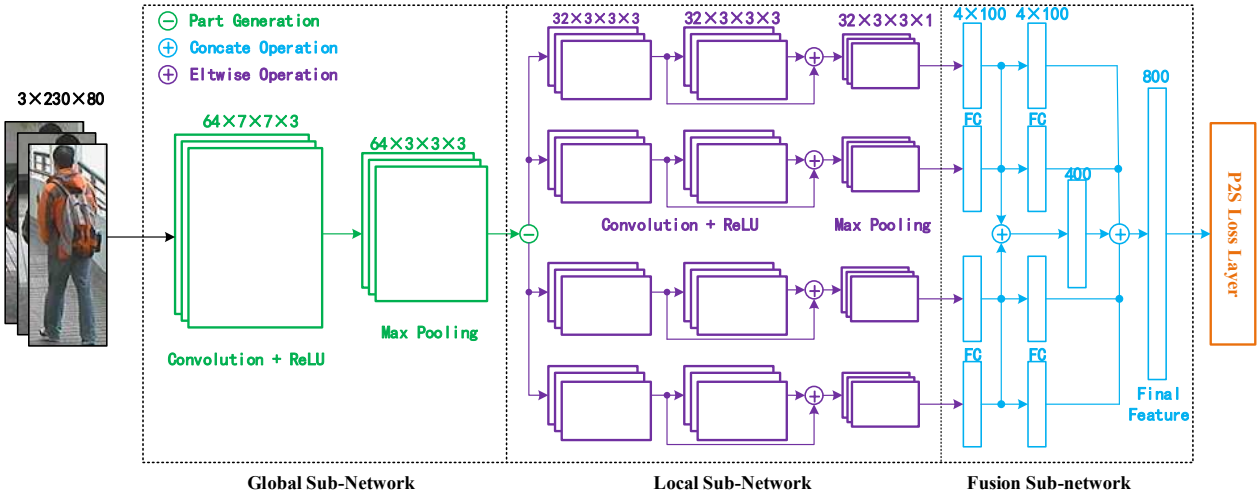

Figure 3. The deep feature learning and fusion neural network. This architecture is comprised of three sub-networks: global sub-network, local sub-network and fusion sub-network. The first two part extract the global feature representations and the local feature representations from person images by using convolutional layer, max-pooling layer and part generation strategy. The third part 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 P2S loss layer for similarity comparison.

64 learned filters of size $7 \times 7 \times 3$. Then, these feature maps are passed through a rectified linear unit (ReLU). Finally, the resulting feature maps are passed through a max pooling kernel of size $3 \times 3 \times 3$ with stride 3.

Local sub-network The second part of our network is a local sub-network, which is consisted of four teams of convolutional layers and max pooling 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. What's more, the outputs of the first local convolutional layer are summarized with the outputs of the second local convolutional layer using eltwise operation. Afterwards, we add a rectified linear unit (ReLU) after them. Finally, the resulting feature maps are passed through max pooling kernels of size 3×3 with stride 1. In order to learn the feature representations of different body parts discriminately, we do not share the parameters among the four teams of convolutional layers.

Fusion sub-network The third part of our network is a fusion sub-network, which is consisted 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 team. The dimension of the fully connected layer is 100 and a rectified linear unit (ReLU) is added between them. Then, the discriminately learned local feature representations of the first four fully connected layers are concated to be summarized by adding another fully connected layers, whose dimension is 400. Finally, the resulting feature representation is further concated with the outputs of the second four fully connected layers to generate 800 dimensional final feature representations. Similarly, we do not share the parameters among the four fully connected layers to keep the discriminative of feature representations of different body parts.

3.3. Optimization

We use the momentum method to update the adaptive weights, and the gradient back-propagation method to optimize the parameters of the deep CNN. Both of them are carried out in a mini-batch pattern. To proceed, we first calculate the gradients of the loss function with respect to both the adaptive weight parameters and the feature representation parameters of the corresponding layer. For simplicity, we consider the parameters in the network as a whole by defining $\mathbf{\Omega}^k = [\mathbf{W}^k, \mathbf{b}^k]$, and $\mathbf{\Omega} = \{\mathbf{\Omega}^1, \ldots, \mathbf{\Omega}^K\}$.

The weight parameters μ, ν can be adaptively learned in the training process by using the momentum method. In order to simplify the problem, we define $\mu = \xi + \vartheta$ and $\nu = \xi - \vartheta$, therefore they can be updated by only updating ϑ. The partial derivative of the triplet term with respect to ϑ can be formulated as follows:

$$t = \begin{cases} \frac{\partial \mathrm{T}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s})}{\partial \vartheta}, & if\ \mathcal{T} > 0\,, \\ 0 & , \ else\,, \end{cases} \tag{10}$$

where $\mathcal{T} = \mathcal{M}_t + \mathrm{T}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s})$, and $\frac{\partial \mathrm{T}}{\partial \vartheta}$ can be computed as follows:

$$\frac{\partial \mathrm{T}}{\partial \vartheta} = 2\mathrm{N}_{i,k}^a [\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{k,s}\|_2^2 - \|\mathbf{x}_B^{j,r} - \mathbf{x}_B^{k,s}\|_2^2]\,, \tag{11}$$

Then ϑ can be updated as follows:

$$\vartheta = \vartheta - \eta \cdot t\,, \tag{12}$$

where η is the updating rate. It can be clearly seen that when $\|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{k,s}\|_2^2 > \|\mathbf{x}_B^{j,r} - \mathbf{x}_B^{k,s}\|_2^2$, namely $t < 0$, then μ will be decreased while ν will be increased; and vice verse. As a result, the strength of back-propagation to each sample in the same triplet unit will be adaptively tuned, in which the anchor and the positive will be clustered, and the negative one will be far away from the hyper-line expanded by the anchor and the positive.

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

$$\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}} = \sum_{i=1}^{N} \mathcal{L}_P(\mathbf{X}_i, \mathbf{\Omega}) + \alpha \mathcal{L}_T(\mathbf{X}_i, \mathbf{\Omega}) + 2\beta \sum_{k=1}^{K} \mathbf{\Omega}^k\,, \tag{13}$$

where the first term represents the gradient of the pairwise term, the second term denotes the gradient of the triplet term, and the third term is the gradient of the regularizer term.

For simplicity, we define $\mathcal{P} = \mathcal{C}_p - \mathrm{G}_{i,j}^a(\mathcal{M}_p - \|\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}\|_2^2)$, then the gradient back-propagation of the pairwise term can be formulated as follows:

$$\mathcal{L}_P = \begin{cases} \frac{\partial \mathbb{P}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r})}{\partial \mathbf{\Omega}}, & if\ \mathcal{P} > 0\,, \\ 0 & , \ else\,, \end{cases} \tag{14}$$

where $\frac{\partial \mathbb{P}}{\partial \mathbf{\Omega}}$ is defined as follows:

$$\frac{\partial \mathbb{P}}{\partial \mathbf{\Omega}} = \frac{1}{\mathcal{Z}_p} \sum_{j=1}^{N} \sum_{k=1}^{M} 2\mathrm{G}_{i,j}^a (\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}) \cdot \frac{\partial \mathbf{x}_A^{i,a} - \partial \mathbf{x}_B^{j,r}}{\partial \mathbf{\Omega}}. \tag{15}$$

By the definition of $\mathcal{T}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s})$ in Eq. (10), we derive the gradient back-propagation of the triplet term as follows:

$$\mathcal{L}_T = \begin{cases} \frac{\partial \mathbb{T}(\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s})}{\partial \mathbf{\Omega}}, & if\ \mathcal{T} > 0\,, \\ 0 & , \ else\,, \end{cases} \tag{16}$$

where $\frac{\partial \mathbb{T}}{\partial \mathbf{\Omega}}$ is defined as follows:

$$\begin{aligned} \frac{\partial \mathbb{T}}{\partial \mathbf{\Omega}} = & \frac{1}{\mathcal{Z}_t} \sum_{j,k=1}^{N} \sum_{r,s=1}^{M} 2\mathrm{P}_{i,j}^a (\mathbf{x}_A^{i,a} - \mathbf{x}_B^{j,r}) \cdot \frac{\partial \mathbf{x}_A^{i,a} - \partial \mathbf{x}_B^{j,r}}{\partial \mathbf{\Omega}} \\ & - 2\mu \mathrm{N}_{i,k}^a (\mathbf{x}_A^{i,a} - \mathbf{x}_B^{k,s}) \cdot \frac{\partial \mathbf{x}_A^{i,a} - \partial \mathbf{x}_B^{k,s}}{\partial \mathbf{\Omega}} \\ & - 2\nu \mathrm{N}_{i,k}^a (\mathbf{x}_B^{j,r} - \mathbf{x}_B^{k,s}) \cdot \frac{\partial \mathbf{x}_B^{j,r} - \partial \mathbf{x}_B^{k,s}}{\partial \mathbf{\Omega}}. \end{aligned} \tag{17}$$

From the above derivations, it is clear that the gradients of both the pairwise term and the triplet term can be easily calculated given the values of $\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}, \mathbf{x}_B^{k,s}$ and $\frac{\partial \mathbf{x}_A^{i,a}}{\partial \mathbf{\Omega}}, \frac{\partial \mathbf{x}_B^{j,r}}{\partial \mathbf{\Omega}}, \frac{\partial \mathbf{x}_B^{k,s}}{\partial \mathbf{\Omega}}$ in each mini-batch, in which they can be obtained by separately running the forward and backward propagation for each image in both the pairwise and triplet units. Because of the algorithm needs to go through all the pairwise and triplet units to accumulate the gradients in each iteration, we call it the P2S gradient descent algorithm. We show the overall process in Algorithm 1.

Algorithm 1 The P2S gradient descent algorithm

Input: Training samples $\mathbf{X}$, learning rate ω, maximum iterations H, weight parameters α and β, initialization to weight parameters μ and ν, updating rate η, margin parameters $\mathcal{C}_p$, $\mathcal{M}_p$ and $\mathcal{M}_t$.
Output: The network parameters $\mathbf{\Omega}$.
repeat
1. Calculate the output feature representations of $\mathbf{x}_A^{i,a}, \mathbf{x}_B^{j,r}$ and $\mathbf{x}_B^{k,s}$ in both the pairwise term and triplet term in a mini-batch by forward propagation.
repeat
a) Update the weight parameters μ and ν according to Eq. (10), Eq. (11) and Eq. (12);
b) Calculate $\frac{\partial \mathbb{P}}{\partial \mathbf{\Omega}}$ and $\frac{\partial \mathbb{T}}{\partial \mathbf{\Omega}}$ according to Eq. (15) and Eq. (17), respectively;
c) Increment the gradient $\frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}}$ according to Eq. (13), Eq. (14) and Eq. (16);
until Traverse all the pairwise and triplet units in each mini-batch.
2. Update $\mathbf{\Omega}_{h+1} = \mathbf{\Omega}_h - \omega_h \frac{\partial \mathrm{L}}{\partial \mathbf{\Omega}_h}$ and $h \leftarrow h + 1$.
until $h > H$

4. Experiments

4.1. Datasets and Settings

Datasets We evaluate our method on four benchmark datasets, namely 3DPeS [2], CUHK01 [17], PRID2011 [12] and Market1501 [42]. Each of them has at least one image for each person and from each camera view.

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

CUHK01: The dataset contains 971 persons captured from two camera views in a campus environment, and there are two images for each person under every camera view. We utilize the same protocol with [31], where 871 person images are used for training and the left for testing.

PRID2011: The dataset includes 749 persons, captured by two disjoint cameras, with sequences lengths of 5 to 675 frames. Following the protocol used in [37], we only consider the first 200 persons, who appear in both cameras.

Market1501: The dataset contains 32668 images of

Table 1. Matching rates(%) on the 3DPeS dataset.

Methods	Top1	Top5	Top10	Top15	Top20
KISSME [15]	22.94	48.71	62.21	72.39	78.11
LF [28]	33.43	45.50	69.98	76.53	81.03
ME [27]	53.30	76.79	86.03	89.37	92.78
kLFDA [35]	54.02	77.74	85.92	90.04	92.38
SCSP [3]	57.29	78.97	85.01	89.52	91.51
Our Method (P2P)	61.97	84.17	92.19	93.85	95.94
Our Method (P2S)	**71.16**	**90.51**	**95.19**	**96.88**	**97.60**

Table 2. Matching rates(%) on the CUHK01 dataset.

Methods	Top1	Top5	Top10	Top15	Top20
KISSME [15]	29.40	59.34	71.45	80.09	88.12
ITML [6]	17.10	41.03	53.12	63.87	69.36
LMNN [32]	21.17	49.49	61.12	69.93	78.32
IDLA [1]	65.00	89.33	92.04	93.74	96.51
JSC [31]	65.71	89.41	92.52	93.74	96.63
Our Method (P2P)	68.91	89.23	94.29	96.35	96.74
Our Method (P2S)	**77.34**	**93.51**	**96.73**	**97.84**	**98.53**

Table 3. Matching rates(%) on the PRID2011 dataset.

Methods	Top1	Top5	Top10	Top20
KISSME [15]	28.54	59.78	72.13	83.26
LF [28]	26.40	56.07	69.89	81.12
LMNN [32]	14.38	38.09	50.22	67.19
LADF [19]	8.20	20.45	29.89	42.25
TDL [37]	30.22	59.10	74.04	88.43
Our Method (P2P)	62.24	88.73	98.61	99.92
Our Method (P2S)	**70.71**	**95.15**	**98.92**	**100.00**

1501 identities. Each identity 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 [38].

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 parameters $\alpha = 0.1, \beta = 0.01$, the direction control parameters $\mu = 0.6, \nu = 0.4$ and the margin parameters $\mathcal{C}_p = 0.2, \mathcal{M}_p = 0.3, \mathcal{M}_t = 1.2$.

Evaluation protocol The dataset is separated into the training set and the testing set, in which images of the same person can only appear in either set. 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 [9], 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.

Comparison Results We compare our results with several existing methods on the four benchmark datasets, namely KISSME [15], LADF [19], LF [28], kLFDA [35], SCSP [3], ITML [6], LMNN [32], ME [27], LDNS [38], JSC [31], TDL [37], Bow [42] and IDLA [1]. In order to analyze how each ingredient contributes to the final performance improvement, we report the results of our method in two variations, *i.e*. P2P and P2S in each of the table, whereas the former P2P results are obtained without the triplet term, and the P2S utilizes the complete constraints. Detailed results are listed from Table 1 to Table 4, where the best performance is highlighted in bold red, and the second best is highlighted in blue.

4.2. Results

Table 1 lists the results on the 3DPeS dataset, in which our P2P method gets the second best performance, contributed by the part-based deep CNN architecture, and our P2S method achieves the best performance in all Top 1 to Top 20 accuracies. Compared with previous best performed method SCSP [3] on this dataset, our two methods outperform it by 4.68% and 13.87% in Top 1 accuracy, respectively. In addition, benefit from the P2S information used in the triplet term, the P2S method wins the P2P method 9.19% in Top 1 accuracy.

The results from the CUHK01 dataset are reported in Table 2, and listed benchmark works include both traditional methods and deep learning based methods. From the results, we can see that our two methods outperform the previous best accuracy achieved by a deep learning based methods IDLA [1] and JSC [31]. In particular, our two methods outperform the JSC method with 3.20% and 11.63% in Top 1 accuracy, respectively. Similar to that in Table 1 , the P2S method beats the P2P method by 8.43% in Top 1 accuracy by taking the P2S information into consideration.

In Table 3, the PRID2011 dataset is specially designed for video based person Re-ID problem. To make fair comparison, we choose to not use any video-based cue in our P2S method, *i.e*. the same way as in [37]. Results again show that, our P2P method wins the second best performance and our P2S method achieves the best performance in all Top 1 to Top 20. Compared with previously the best method TDL [37], our two proposed methods outperform it by 32.02% and 40.49% in Top 1 accuracy, respectively. In addition, the P2S method wins the P2P method by 8.47% 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 [38]. As illustrated in Table 4, the proposed two methods outperform LDNS by 0.29% and 9.70% in Top 1 accuracy under the single-query setting, and 4.53% and 14.22% in Top 1 accuracy under the multi-query setting, respectively. Again, the presented P2S method wins the P2P method by 9.41% and 9.69% 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.

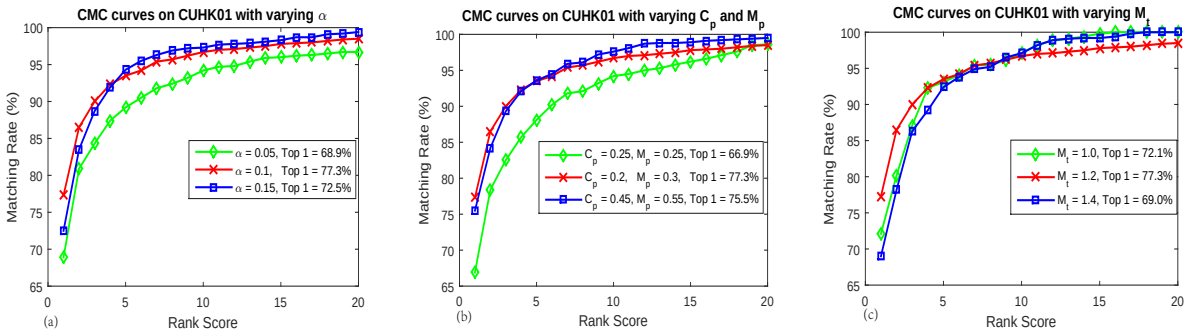

Figure 4. CMC curves on the CUHK01 dataset with varying parameters, in which (a) shows the matching results with varying α and setting $\mathcal{C}_p = 0.2, \mathcal{M}_p = 0.3$ and $\mathcal{M}_t = 1.2$, (b) shows the matching results with varying $\mathcal{C}_p, \mathcal{M}_p$ and setting $\alpha = 0.1$ and $\mathcal{M}_t = 1.2$; and (c) shows the matching results with varying $\mathcal{M}_t$ and setting $\alpha = 0.1, \mathcal{C}_p = 0.2, \mathcal{M}_p = 0.3$.

Table 4. Matching rates(%) on the Market1501 dataset.

Methods	Single-Query		Multi-Query	
	Top1	mAP	Top1	mAP
Bow [42]	34.38	14.10	42.64	19.47
kLFDA [35]	51.37	24.43	52.67	27.36
KISSME [15]	40.50	19.02	−−	−−
LDNS [38]	61.02	35.68	71.56	46.03
SCSP [3]	51.90	26.35	−−	−−
Our Method (P2P)	61.31	35.71	76.09	47.92
Our Method (P2S)	**70.72**	**44.27**	**85.78**	**55.73**

Table 5. Influence of the direction control parameters.

Datasets	μ=1.0, ν=0.0		μ=0.6, ν=0.4		μ=0.4, ν=0.6	
	Top1	Top5	Top1	Top5	Top1	Top5
3DPeS	64.38	88.10	**71.16**	**90.51**	69.28	89.91
CUHK01	71.27	92.02	**77.34**	**93.51**	74.54	92.83
PRID2011	65.32	90.19	**70.71**	**95.15**	68.24	93.32
Market1501	63.82	88.78	**70.72**	**90.52**	68.21	89.09

Parameter Analysis As observed in our experiments, the weight parameter α, the margin parameters $\mathcal{C}_p, \mathcal{M}_p, \mathcal{M}_t$, and the direction control parameters μ, ν have major effects to our method. In the following, we give an empirical analysis of our method on the CUHK01 dataset.

The influence of parameters $\mathcal{C}_p, \mathcal{M}_p, \mathcal{M}_t$ and α is shown in Fig. 4, in which we analyze the influence by changing one parameter while fixing the others. From the results, we can see that our method achieves its best performance by setting $\alpha = 0.1, \mathcal{C}_p = 0.2, \mathcal{M}_p = 0.3$ and $\mathcal{M}_t = 1.2$. Besides, we can conclude the following three empirical conclusions: 1) For parameter α, large value will lead to the over-fitting problem and small value will weaken the strength of P2S constraint. 2) For parameters $\mathcal{C}_p, \mathcal{M}_p$, small down-margin will lead to the over-fitting problem, and large upper-margin will make the numerical instability. 3) Similarly, large $\mathcal{M}_t$ will also lead to the numerical instability and small $\mathcal{M}_t$ will make the candidate positive and negative samples undistinguishable.

Different from the conventional triplet formulation proposed by [7], our symmetric triplet framework introduces a weighed negative distance term to optimize the back-propagation of each sample in one triplet unit. Therefore, the conventional triplet formulation is a special case of our method by setting $\mu = 1.0, \nu = 0.0$ and $\eta = 0.0$. The comparison results are shown in Table 5, in which our symmetric triplet framework outperforms the conventional one with $6.78\%, 6.07\%, 5.39\%$ and 6.90% in Top 1 on the four datasets, respectively. Benefit from the parameter updating strategy, the initial values of μ and ν may have a slight effect to our method, in which we can see the performance only fall $1.88\%, 2.80\%, 2.47\%$ and 2.51% by setting $\mu = 0.4, \nu = 0.6$ on the four datasets, respectively.

5. Conclusion

In this paper, we propose a novel person re-identification method by point to set (P2S) similarity comparison in a part-based deep CNN to perform integrated feature learning and fusion. The deep architecture learns the global features, local features and fused features in the global sub-network, local sub-network and fusion sub-network, respectively. The P2S distance metric jointly minimizes the intra-class distance and maximizes the inter-class distance, while back-propagating the gradient to optimize the deep parameters. As a result, the learned deep ranking model can effective distinguish different persons by learning discriminative and stable features. Experiment results on the 3DPeS, CUHK01, PRID2011 and Market1501 datasets show that our method outperforms the state-of-the-art approaches in person re-identification.

Acknowledgement

This work is partially supported by National Basic Research Program of China (973 Program) under Grant No. 2015CB351705, and the National Science Foundation of China under Grant No. 61473219.

References

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

[2] D. Baltieri, R. Vezzani, and R. Cucchiara. Sarc3d: a new 3d body model for people tracking and re-identification. In *Proceedings of the 16th International Conference on Image Analysis and Processing*, pages 197–206, Ravenna, Italy, Sept. 2011.

[3] D. Chen, Z. Yuan, B. Chen, and N. Zheng. Similarity learning with spatial constraints for person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[4] D. Chen, Z. Yuan, G. Hua, N. Zheng, and 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*, pages 1565–1573, 2015.

[5] D. S. Cheng, M. Cristani, M. Stoppa, L. Bazzani, and V. Murino. Custom pictorial structures for re-identification. In *BMVC*, volume 1, page 6. Citeseer, 2011.

[6] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216. ACM, 2007.

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

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

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

[10] D. Gray and H. Tao. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In *European conference on computer vision*, pages 262–275. Springer, 2008.

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] M. Hirzer, C. Beleznai, P. M. Roth, and H. Bischof. Person re-identification by descriptive and discriminative classification. In *Scandinavian conference on Image analysis*, pages 91–102. Springer, 2011.

[13] M. Hirzer, P. M. Roth, M. Köstinger, and H. Bischof. Relaxed pairwise learned metric for person re-identification. In *Computer Vision–ECCV 2012*, pages 780–793. Springer, 2012.

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

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

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[17] W. Li and X. Wang. Locally aligned feature transforms across views. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3594–3601. IEEE, 2013.

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

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

[20] S. Liao, Y. Hu, X. Zhu, and 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)*, June 2015.

[21] C. C. Loy, C. Liu, and S. Gong. Person re-identification by manifold ranking. In *Image Processing (ICIP), 2013 20th IEEE International Conference on*, pages 3567–3571. IEEE, 2013.

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

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

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

[25] B. T. Morris and M. M. Trivedi. A survey of vision-based trajectory learning and analysis for surveillance. *Circuits and Systems for Video Technology, IEEE Transactions on*, 18(8):1114–1127, 2008.

[26] H. V. Nguyen and L. Bai. Cosine similarity metric learning for face verification. In *Computer Vision–ACCV 2010*, pages 709–720. Springer, 2011.

[27] S. Paisitkriangkrai, C. Shen, and 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*, pages 1846–1855, 2015.

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

[29] B. Prosser, W.-S. Zheng, S. Gong, T. Xiang, and Q. Mary. Person re-identification by support vector ranking. In *BMVC*, volume 2, page 6, 2010.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] F. Wang, W. Zuo, L. Lin, D. Zhang, and 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)*, June 2016.

[32] K. Q. Weinberger, J. Blitzer, and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. In *Advances in neural information processing systems*, pages 1473–1480, 2005.

[33] Z. Wu, Y. Li, and 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):1095–1108, 2015.

[34] T. Xiao, H. Li, W. Ouyang, and 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)*, June 2016.

[35] F. Xiong, M. Gou, O. Camps, and M. Sznaier. Person re-identification using kernel-based metric learning methods. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[36] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Deep metric learning for person re-identification. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 34–39. IEEE, 2014.

[37] J. You, A. Wu, X. Li, and W.-S. Zheng. Top-push video-based person re-identification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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

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

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

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

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

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

[44] S. Zhou, J. Wang, Q. Hou, and Y. Gong. Deep ranking model for person re-identification with pairwise similarity comparison. In *Pacific Rim Conference on Multimedia*, pages 84–94. Springer, 2016.