Multimed Tools Appl (2018) 77:3049–3069
DOI 10.1007/s11042-017-5009-y

A loss combination based deep model for person re-identification

Fuqing Zhu[1] · Xiangwei Kong[1] · Qun Wu[2,3] · Haiyan Fu[1] · Ming Li[1]

Received: 14 March 2017 / Revised: 28 June 2017 / Accepted: 4 July 2017 /
Published online: 8 July 2017

Abstract The Convolutional Neural Network (CNN) has significantly improved the state-of-the-art in person re-identification (re-ID). In the existing available identification CNN model, the softmax loss function is employed as the supervision signal to train the CNN model. However, the softmax loss only encourages the separability of the learned deep features between different identities. The distinguishing intra-class variations have not been considered during the training process of CNN model. In order to minimize the intra-class variations and then improve the discriminative ability of CNN model, this paper combines a new supervision signal with original softmax loss for person re-ID. Specifically, during the training process, a center of deep features is learned for each pedestrian identity and the deep features are subtracted from the corresponding identity centers, simultaneously. So that, the deep features of the same identity to the center will be pulled efficiently. With the combination of loss functions, the inter-class dispersion and intra-class aggregation

✉ Xiangwei Kong
kongxw@dlut.edu.cn

Fuqing Zhu
fuqingzhu@mail.dlut.edu.cn

Qun Wu
wuq@zstu.edu.cn

Haiyan Fu
fuhy@dlut.edu.cn

Ming Li
mli@dlut.edu.cn

[1] School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China

[2] General Design Institute, Zhejiang Sci-Tech University, Hangzhou 310018, China

[3] Taizhou Research Institute, Zhejiang University, Taizhou 318000, China

can be constrained as much as possible. In this way, a more discriminative CNN model, which has two key learning objectives, can be learned to extract deep features for person re-ID task. We evaluate our method in two identification CNN models (i.e., CaffeNet and ResNet-50). It is encouraging to see that our method has a stable improvement compared with the baseline and yields a competitive performance to the state-of-the-art person re-ID methods on three important person re-ID benchmarks (i.e., Market-1501, CUHK03 and MARS).

Keywords Convolutional neural network · Loss combination · Person re-identification

1 Introduction

Person re-identification (re-ID) is the task of matching people under cross scenarios, which has received increasing attention in automated surveillance system for its potential applications in human retrieval, cross-camera tracking and so on. From the perspective of computer vision community, the greatest challenge in person re-ID is how to match two images of the same person correctly under the variances of lighting, pose and viewpoint. Person re-ID lies in between image classification [6, 54, 57] and retrieval [55, 56, 62, 64, 66, 68]. In the task of image classification, the training samples are available for each class, and the testing samples fall into these pre-defined classes. In the task of image retrieval, there is no training sample because the content of query is unknown in advance while the gallery may contain various types of objects. Therefore, the training classes are not available and the testing classes are previously unseen. Compared to the image classification task, person re-ID is similar in that the training classes are available, which includes available samples of different identities. While, person re-ID is also similar to image retrieval in that the testing classes are previously unseen, that is there is no overlap between the training and testing identities. Previous person re-ID works [29, 30, 63, 65] usually take advantage of both image classification and retrieval. Despite the best efforts from the researchers, the person re-ID is still an unsolved problem and has widespread practical application values.

The previous mainstream works in person re-ID typically focus on visual feature representation [2, 21, 50, 61] and distance metric learning [26, 35, 60]. Recently, deep learning has achieved record leading in person re-ID community [43, 44, 51, 65]. Among the deep learning based methods for person re-ID, in an attempt to make full use of pedestrian identity annotation, the identification CNN model is employed to learn deep feature for each pedestrian image. On the large-scale datasets (e.g., PRW [69] and MARS [65]), the identification CNN model has achieved good re-ID performance without any special training sample selection process. Yet the application of this identification CNN model requires a rich amount of training samples per identity for model convergence. In the existing available identification CNN model, the softmax loss function is employed as the supervision signal to train the CNN model. The softmax loss only encourages the separability of learned deep features between different identities. However, a concurrent problem is that the intra-class variances have not been considered during the training process of CNN model. In the person re-ID datasets, the existing training samples exhibit obviously intra-class variances as shown in Fig. 1 under large variances of lighting, pose and viewpoint. The neglect of intra-class variances in training set may compromise the discriminative ability of CNN model. In this paper, we will thus provide a possible solution on how to improve the discriminative ability of CNN model for person re-ID in which the inter-class dispersion and intra-class aggregation can be constrained as much as possible.

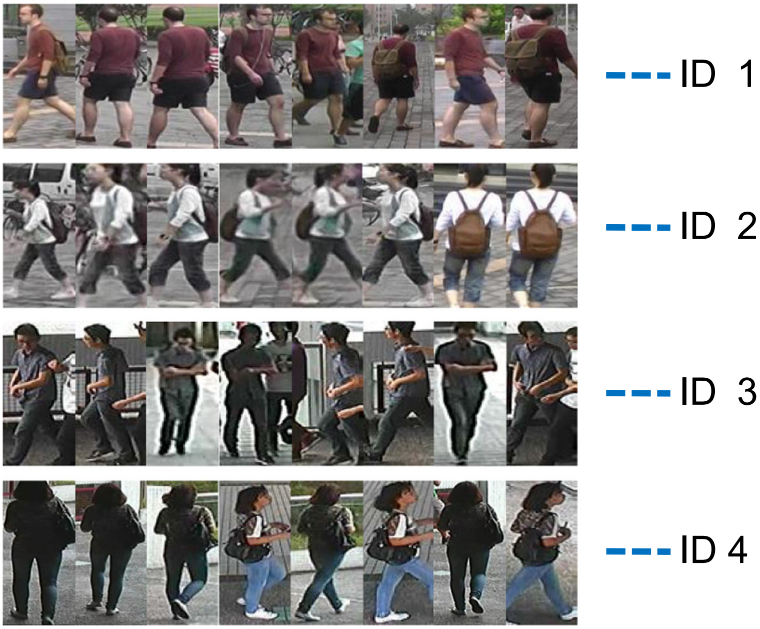

Fig. 1 Training samples of some identities in Market-1501 [63] (*top two rows*) and CUHK03 [29] (*bottom two rows*) datasets. We observe that the intra-class variances are obvious

Given the above considerations, this paper is motivated by the strengths of the two loss functions (i.e., softmax loss and center loss) and leverages their complementary aspects to improve the discriminative ability of CNN model for person re-ID. During the training process of CNN model, the inter-class and intra-class variances are both constrained, simultaneously. To summarize, the main contributions of this paper are listed below.

- We combine the softmax loss and center loss functions to further minimize the intra-class variances for training a more discriminative CNN model for person re-ID.
- On three important person re-ID benchmarks (i.e., Market-1501 [63], CUHK03 [29] and MARS [65]), our loss combination identification CNN model demonstrates the consistent improvement over the corresponding baseline and yields a very competitive re-ID performance compared with state-of-the-art person re-ID methods.

The rest of the paper is organized as follows. In Section 2, we review the related work briefly. Our loss combination based identification CNN model for person re-ID is described in Section 3. In Section 4, extensive experimental results are presented on Market-1501, CUHK03 and MARS datasets. Finally, we conclude this paper in Section 5.

2 Related work

This paper focuses on improving the discriminative ability of the existing identification CNN model for person re-ID. In this section, we will discuss the related works in image-based and video-based person re-ID based on deep learning.

2.1 Image-based person re-ID based on deep learning

Deep learning has been popular in computer vision community since Krizhevsky et al. [25] won ILSVRC 2012. The success of deep learning spreads to person re-ID community in 2014, when the two person re-ID works [29, 58] employ deep learning to determine whether a pair of input pedestrian images belong to the same identity. Generally, there are two types of CNN models that are commonly employed in person re-ID. The first type is the identification CNN model which has been widely used in image classification [25] and object detection [17]. The second type is the siamese model which employs pedestrian image pairs or triplets as input samples and is used in image retrieval [36] and face recognition [39].

In person re-ID community, the major bottleneck of deep learning based method is the lack of training samples for each identity. Because of the difficulty in pedestrian annotation, most of the person re-ID datasets provide only a few images for each identity such as VIPeR [18], iLIDS [70], CUHK01 [28], CUHK02 [27], RAiD [11] and PRID 450S [38]. Currently, most of the deep learning based person re-ID methods focus on the Siamese model due to the number of training samples for each identity is limited. Yi et al. [58] employ a Siamese model, in which the image is first partitioned into three overlapping horizontal parts, and then the parts go through two convolutional layers and finally are fused by a fully connected layer. The filter pairing network architecture designed in [29] is that a patch matching layer is imposed which evaluates the convolutional layer responses of two images in different horizontal stripes. Ahmed et al. [1] design an improved Siamese CNN architecture by imposing a special layer to learn the cross-image representation via computing the neighborhood distance between two input images. Wang et al. [45] design a joint learning Siamese CNN framework, in which the matching of single-image representation and the classification of cross-image representation are jointly learned for pursuing better re-ID accuracy with moderate computational cost. The combination of two representations is utilized the advantages of respective properties that the single-image representation is efficient in matching, while cross-image representation is effective in modeling the relationship between probe and gallery images. Wu et al. [49] design a CNN architecture called "PersonNet", in which the network depth is increased using convolutional filters of smaller sizes. Varior et al. [44] integrate the long short-term memory (LSTM) module into the Siamese model. The multiple LSTM modules process image regions sequentially for memorizing the spatial connections of regions to enhance the discriminative ability of the learned deep features. Varior et al. [43] propose to insert a gating function after each convolutional layer in the Siamese model to capture effective subtle patterns. However, the Siamese model based re-ID method has disadvantage in time efficiency (especially in large-scale datasets). The queries have to pair with each gallery image before being sent into the network during testing, although it achieves state-of-the-art performance on several re-ID benchmarks. While the above works use image pairs as input, Cheng et al. [10] design a triplet based Siamese CNN architecture which employs triplet images as input. The four overlapping body parts for each pedestrian are partitioned after the first convolutional layer, and fused with a global one in the fully connected layer. Liu et al. [33] propose a multi-scale triplet-based CNN architecture which integrates deep and shadow networks to capture pedestrian appearance at different scales. Su et al. [40] propose a Siamese CNN architecture which employs an attribute based triplet loss trained on datasets with identities.

With the scale growth of person re-ID datasets, each pedestrian has a rich amount training samples. Xiao et al. [51] train identities from multiple datasets directly with the identification CNN model. An impact score is proposed for each fully connected neuron, while a domain-guided dropout is imposed based on the impact score. The learned deep features yield excellent re-ID performance. Zheng et al. employ the identification CNN model to achieve excellent re-ID performance on larger datasets PRW [69] without careful training sample selection process. Xiao et al. [52] train a identification model to jointly consider the pedestrian detection and the person re-ID problem on a large-scale dataset which contains 99,809 annotated bounding boxes for 8,432 pedestrian identities. Zheng et al. [71] introduce the unlabeled samples generated by generative adversarial network (GAN) [37] to improve the re-ID performance of the identification CNN model. Lin et al. [31] propose an attribute-person recognition (APR) network based on the identification CNN model, which learns a identity embedding and predicts the pedestrian attributes simultaneously. Sun et al. [41] propose to decorrelate the learned weight vectors using singular vector decomposition (SVD) based on the identification CNN model. In the person re-ID survey [67], some baseline results are presented for both the Siamese and identification CNN models on Market-1501 [63] dataset. From which, the identification CNN model outperforms Siamese model. Because a main drawback of the Siamese model is that it does not make full use of pedestrian identity annotations. The contrastive loss or triplet loss layer in the Siamese model is formulated the similarity of pair-wise or triplet samples, which is the weekly supervision representation of constraining an image pair is belong to the same identity or not. Moreover, as the scale growth of dataset, the number of pair-wise or triplet based samples will grow exponentially. It will bring a convergence problem during the training process of CNN model. Overall, the identification CNN model is more suitable for practical application in large-scale person re-ID.

2.2 Video-based person re-ID based on deep learning

In video-based person re-ID, the data volume of dataset is typically larger since each tracklet contains a number of image sequences such as ETHZ [15], 3DPES [3], PRID 2011 [20], iLIDS-VID [46] and MARS [65] datasets. Since most of these image sequences for each tracklet contain pedestrians, we could use these image sequences directly for training CNN model rather than employing complex video semantic pooling [8] operation. Meanwhile, the large data volume is beneficial for training a scalable CNN model. In [65], the identification CNN model is adopted to train identities, and has been proved the effectiveness of loss convergence and performance improvement. In [53], a novel recurrent feature aggregation framework is proposed to exploit a globally discriminative feature representation from a sequence of tracked person patches by long short term memory (LSTM). Compared with image-based person re-ID, the video is regarded as multiple image sequences for each matching unit. So the video matching strategy is employed by multi-match or a single-match after video pooling, respectively. The multi-match strategy leads to higher computational cost, especially become problematic on large-scale datasets. The single-match strategy has better scalability which aggregates the frame-level features into a global feature representation. In this work, we follow the single-match strategy after video pooling (average and max pooling) adopted in [65] for video-based person re-ID on MARS dataset.

Recently, a center loss function [48] is designed to minimize intra-class variances for robust face recognition. Compared with the task of face recognition, the phenomenon of intra-class variances is also obvious in person re-ID. But there are significant differences between the two tasks. Our work departs from previous methods in two aspects. First, this work focuses on person re-ID, a task in which the intra-class variances is obvious and the number of training samples for each identity is much fewer. Second, the center loss is imposed on the conventional identification CNN model to further constrain the intra-class

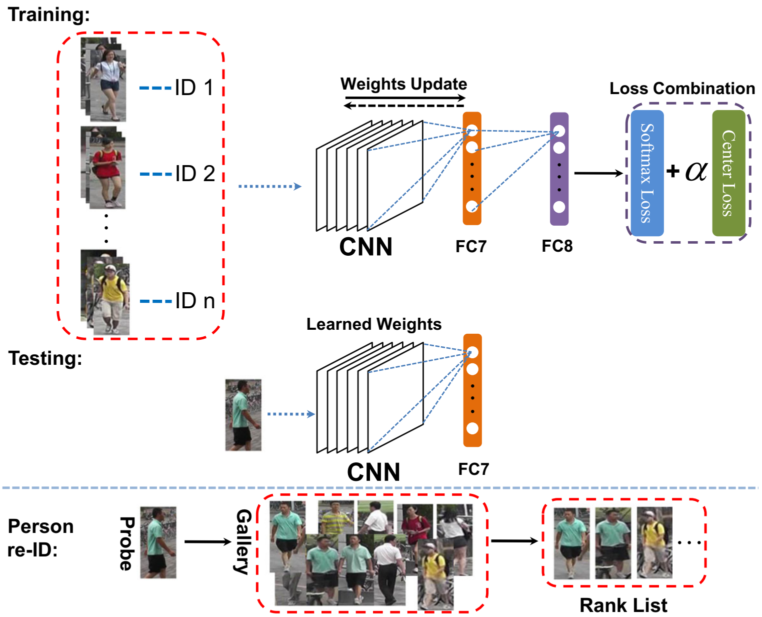

Fig. 2 The overall architecture of our loss combination method based on the identification CNN model for person re-ID

features aggregation and keep inter-class features dispersion at the same time. The center loss and softmax loss have complementary advantages. The combination of center loss and softmax loss is proved to be an effectiveness way in the person re-ID task.

3 Our approach

In this section, we elaborate our loss combination method based on the identification CNN model for person re-ID. We first describe the overall architecture of deep neural network, which is an ID-discriminative Embedding deep learning framework for each identity. Then, we show the details of our method including the loss function combination, the training process of CNN model, and the testing process of person re-ID. Figure 2 illustrates the overall architecture of our loss combination method based on the identification CNN model[1] for person re-ID.

3.1 Overall architecture of deep neural network

Our network is an identification CNN model in which the softmax loss and center loss are combined to encourage the learned deep features of inter-class dispersed separably and intra-class aggregated compactly, simultaneously. In this way, the differences of the inter-class features are enlarged, while the intra-class features variances are reduced. So that, the

[1] Here, we do not provide detailed descriptions of the identification CNN model and just take CaffeNet [25] model as instance in Fig. 2.

discriminative ability of CNN model can be significantly improved for person re-ID. In this paper, the learned deep features will be directly used for person re-ID without any additional feature selection [5, 7] process in a simple way. The training process of the identification CNN model is essentially learning an ID-discriminative embedding in the person subspace for each identity. The testing process of person re-ID can be regarded as a special image search task, in which the deep features of pedestrian images in probe and gallery sets are extracted by the trained CNN model.

The training set is given as $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{N}$, in which the pedestrian image is $\mathbf{x}_i$, and the ID is y_i. The goal is to train a identification CNN model $\mathbb{M}$, which can be regarded as a mapping $\mathrm{f}(\mathbf{x}, \theta) \in \mathbb{R}^C$, where θ represents the parameters of each layer in the network. The θ is updated using optimization algorithm in each mini-batch iteration. The t-th iteration updates the current parameters θ_t as following formula:

$$\theta_{t+1} = \theta_t + \gamma \cdot \frac{1}{|\mathcal{D}_t|} \sum_{(\mathbf{x},\mathbf{y} \in \mathcal{D}_t)} \nabla_{\theta t} \left[l\left(\mathbf{x},\mathbf{y}\right) \right] \tag{1}$$

where γ is learning rate, $\mathcal{D}_t$ is a mini-batch randomly selected from the training set $\mathcal{D}$, ∇ is gradient operation and l is the loss function.

During the training process of CNN model, 10% of the samples in training set are randomly selected as validation samples. The settings of intermediate layers (i.e., convolutional layer, pooling layer and fully connected layer) are default. The output number of last fully connected layer (FC8 in Fig. 2) is setting as the total number of identities in the training set (i.e., 751, 1,160 and 631 on Market-1501 [63], CUHK03 [29] and MARS [65] datasets, respectively). The loss layer is the supervision signal that guides the training process of the identification CNN model, while the convergence of deep neural network should be guaranteed. The identification CNN model is fine-tuned from the ImageNet [14] pre-trained model.

3.2 Loss function combination

The softmax loss function is employed to force the learned deep features of inter-class staying apart, while the center loss function is designed to pull the learned deep features of intra-class to the center. The two loss functions have complementary advantages to train a more discriminative identification CNN model for person re-ID.

The softmax loss function is:

$$\mathcal{L}_s = -\sum_{i=1}^{m} \log \frac{e^{W_{y_i}^T \mathbf{x}_i + b_{y_i}}}{\sum\limits_{j=1}^{C} e^{W_j^T \mathbf{x}_i + b_j}}, \tag{2}$$

where W_j denotes the j-th column of the weights W in the last fully connected layer, and b is the bias term. The size of mini-batch is m. The number of identities is C. Under the supervision of softmax loss between last fully connected layer response and ground-truth ID, the learned deep features are separable for each identity. Since the intra-class variances are significant on re-ID datasets, the learned deep features are not discriminative enough. Therefore, it is not optimal to use the learned deep features directly for person re-ID task. The center loss function is an effective way to improve the discriminative ability of the learned deep features.

The center loss function is formulated as follows:

$$\mathcal{L}_c = \sum_{i=1}^{m} \left\| \mathbf{x}_i - \mathbf{c}_{y_i} \right\|_2^2, \tag{3}$$

where $\mathbf{c}_{y_i}$ denotes the y_i-th class center of deep features. The $\mathbf{c}_{y_i}$ is updated based on each mini-batch. In each mini-batch iteration, the centers are calculated by averaging the features of the corresponding identities. When the $t = y_i$ is satisfied, the update equation of center $\mathbf{c}_{y_i}$ is as follows:

$$\Delta c_t = \frac{\sum_{i=1}^{m} c_t - \mathbf{x}_i}{1+m}, \tag{4}$$

where m is the size of mini-batch.

We combine the softmax loss $\mathcal{L}_s$ and center loss $\mathcal{L}_c$ to train the identification CNN for a more discriminative feature learning. The final loss function is

$$\mathcal{L} = \mathcal{L}_s + \alpha \mathcal{L}_c, \tag{5}$$

where the parameter α is assigned to 0.008 in our experiment.

The above two loss formulations effectively deal with the variances of inter-class and intra-class. They are complementary and jointly optimize the identification CNN model effectively. If we only supervise the CNN training by the softmax loss function, the learned deep features for each identity contain large intra-class variances. Meanwhile, if we only use the center loss function as the supervision signal during the CNN model training, the learned deep features and centers for each identity would decrease to zeros. Using either of them could not obtain a optimal CNN model for person re-ID. Intuitively, minimizing the intra-class variations and keeping the features of different identities separable are both significant for training a more discriminative identification CNN model.

3.3 Details of training and testing

Details of training The training pedestrian images are first resized to 256×256, then randomly cropped into a fixed size[2] with random horizontal flipping. The training samples are shuffled before fed into the data layer of CNN model for each iteration. The mean image is subtracted from all training images. For video-based re-ID dataset (i.e., MARS), an input tracklet is decomposed into some image sequences. The training set is composed of image sequences in video-based re-ID dataset.

We use the CAFFE [22] package to implement our method with two CNN models: CaffeNet [25] and ResNet-50 [19], respectively. We adopt the mini-batch stochastic gradient descent (SGD) [4] to update the parameters of each layer. The baseline is that only softmax loss function (when $\alpha = 0$) is employed as supervision signal for training the CNN model.

Details of testing For any pedestrian image, we feed forward the image to the trained identification CNN model $\mathbb{M}$. The learned deep feature of the pedestrian image is extracted from the fully connected layer or pooling layer.[3] In this way, the deep features of pedestrian images in the probe (query) and gallery (database) sets are obtained. The similarity can be calculated by Euclidean distance of corresponding deep features. By sorting the distance between samples in two sets (probe and gallery), the final person re-ID result could be obtained based on rank list.

[2] The size is 227×227 for CaffeNet [25], while is 224×224 for ResNet-50 [19].

[3] Note: CaffeNet [25] is FC7 layer, while ResNet-50 [19] is Pool5 layer.

For video-based re-ID dataset, we feed forward the image sequences of any pedestrian tracklet to the trained CNN model $\mathbb{M}$. The learned deep features of image sequences can be extracted from the immediate layer as same as the image-based re-ID. For each tracklet, we can obtain a series of the learned deep features. And then, max or average pooling is employed to generate a single vector representation. Next, the re-ID process of video-based is the same as image-based method.

4 Experiments

We evaluate our method on three person re-ID benchmarks (i.e., Market-1501 [63], CUHK03 [29] and MARS [65]), in which the first two datasets are image-based and the third is video-based, respectively. The three re-ID datasets are closer towards realistic situations than previous ones. The baseline is the identification CNN model, which only employs the softmax loss as supervision signal for person re-ID. We first show the comparison results of the baseline and our method, and then we provide some comparisons with the state-of-the-art person re-ID methods on three public benchmarks to demonstrate the effectiveness of our method.

We adopt the Cumulated Matching Characteristics (CMC) curve and mean Average Precision (mAP) for evaluation. The CMC shows the probability that a query identity appears in the ranking lists of different sizes. This evaluation protocol is generally believed to focus on precision. In case of there is only one ground-truth match for a given query, the precision and recall are the same issue. However, if multiple ground-truths exist, the CMC is biased because recall is not considered. For the above three benchmarks, there are several cross-camera ground-truths for each query. The mean Average Precision (mAP) provides a more comprehensive evaluation for person re-ID, in which both the precision and recall are considered in the rank list. In the experiments, we both report the results of CMC (The **rank-1** accuracy is shown when CMC curve is absent.) and mAP for the baseline and our method on the three testing datasets.

4.1 Experiments on market-1501 dataset

The Market-1501 [63] dataset contains 32,668 bounding boxes of 1,501 identities. The generation of bounding boxes is automatically labeled by the pedestrian detector DPM [16] completely. The total 1,501 identities are split into 751 identities for training and 750 identities for testing, followed by the protocol in [63]. The initial learning rate is set to 0.001 and reduced by a factor of 0.1 after each 15 epochs. Training is done after iteration 75 epochs. The testing process is performed in the cross-camera mode.

We evaluate our method based on two identification CNN models (i.e., CaffeNet [25] and ResNet-50 [19]), with a comparison with the baseline on Market-1501 dataset in Table 1. We observe from Table 1 that the rank-1 accuracy increases from 56.03 to 56.26% (+0.23%), and an obviously improvement can be seen from mAP, from 32.41 to 34.67% (+2.26%) on CaffeNet model. On ResNet-50 model, the rank-1 accuracy increases from 72.54 to 73.07% (+0.53%), the mAP increases from 46.00 to 51.35% (+5.35%). From the results of the two CNN models, we observe that an even larger improvement can be seen from mAP than rank-1 accuracy on Market-1501 dataset. The mAP provides a more comprehensive evaluation for person re-ID when multiple ground-truths exist, in which both the precision and recall are considered in the rank list. The experiment results illustrate the effectiveness of our method. Figures 3 and 4 summarize the CMC of the baseline and our method comparison on the two CNN models, respectively. The CMC has a stable improvement from rank-1 to rank-10 on CaffeNet and ResNet-50 models.

Table 1 Comparison of the baseline and our method (rank-1 accuracy (%) and mAP (%)) employing different CNN models on Market-1501 dataset

Methods	CaffeNet [25]		ResNet-50 [19]	
	rank-1	mAP	rank-1	mAP
Baseline	56.03	32.41	72.54	46.00
Ours	**56.26**	**34.67**	**73.07**	**51.35**

The experimental results of our method are highlighted in bold

Comparison with the state-of-the-art re-ID methods We first compare our method with the conventional non-deep learning method (i.e., feature design followed by distance metric learning). The feature includes the BoW [63] descriptor and LOMO [30]. The distance metric learning methods include LMNN [47], ITML [12], KISSME [24] and XQDA [30]. As can be seen in Table 2, compared with the conventional non-deep learning method, our method brings significant improvement over benchmark in both rank-1 accuracy and mAP on Market-1501 dataset. Then we compare with some state-of-the-art person re-ID methods based on deep learning, including PersonNet [49], Semi-supervised Deep Attribute Learning (SSDAL) [40], WARCA [23], Temporal Model Adaptation (TMA) [34], End-to-end Comparative Attention Network (CAN) [32], Multi-region Bilinear CNN [42], SCSP [9], Null Space [59], Siamese LSTM [44] and Gated S-CNN [43]. From the results in Table 2, it is clear that our method significantly outperforms most of deep learning methods in both rank-1 accuracy and mAP on Market-1501 dataset. Our method is an identification CNN model, which makes full use of pedestrian ID annotation information and encourages

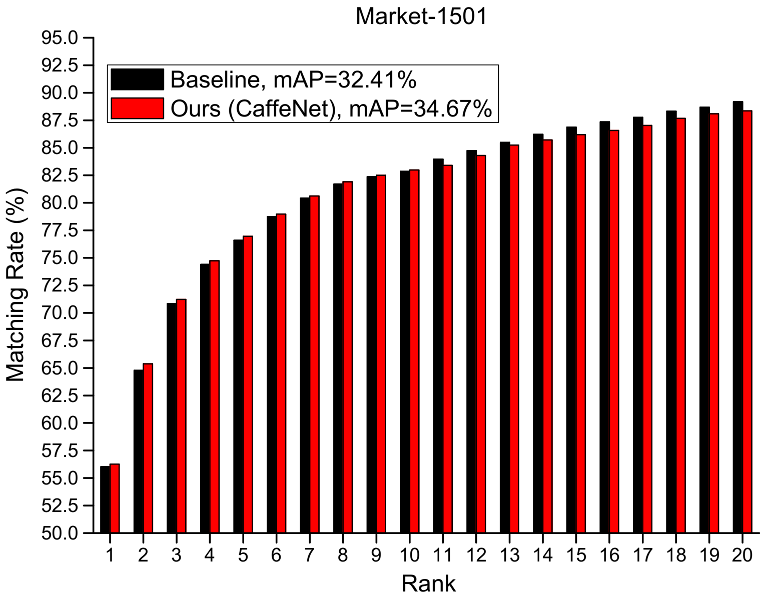

Fig. 3 The CMC of the baseline and our method comparison employing CaffeNet [25] model on Market-1501 dataset

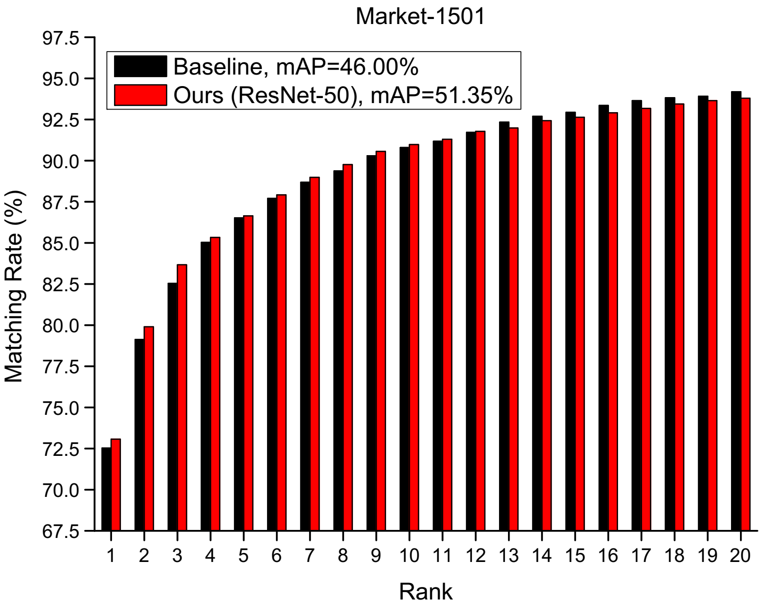

Fig. 4 The CMC of the baseline and our method comparison employing ResNet-50 [19] model on Market-1501 dataset

the learned deep features of inter-class dispersed separably and intra-class aggregated compactly simultaneously. The experimental results demonstrate the effectiveness of our method that the softmax loss and center loss are jointly employed as supervision signal for person re-ID task.

Table 2 Comparison with the state-of-the-art person re-ID methods (rank-1 accuracy (%) and mAP (%)) on Market-1501 dataset

Methods	Market-1501	
	rank-1	mAP
BoW+HS [63]	47.25	21.88
BoW+LMNN [47]	34.00	15.66
BoW+ITML [12]	38.21	17.05
BoW+KISSME [24]	39.61	17.73
LOMO+XQDA [30]	26.07	7.75
PersonNet [49]	37.21	18.57
SSDAL [40]	39.4	19.6
WARCA [23]	45.16	–
TMA [34]	47.92	22.31
End-to-end CAN [32]	48.24	24.43
Multiregion Bilinear CNN [42]	45.58	26.11
SCSP [9]	51.90	26.35
Null Space [59]	55.43	29.87
Siamese LSTM [44]	61.60	35.30
Gated S-CNN [43]	65.88	39.55
Ours (CaffeNet [25])	56.26	34.67
Ours (ResNet-50 [19])	**73.07**	**51.35**

The experimental results of our method are highlighted in bold

Table 3 Comparison of baseline and our method (rank-1 accuracy (%) and mAP (%)) employing different CNN models on CUHK03 dataset

Methods	CaffeNet [25]		ResNet-50 [19]	
	rank-1	mAP	rank-1	mAP
Baseline	53.90	60.11	54.50	60.72
Ours	**54.40**	**60.60**	**56.65**	**61.97**

The experimental results of our method are highlighted in bold

4.2 Experiments on CUHK03 dataset

The CUHK03 [29] dataset contains 13,164 bounding boxes of 1,360 identities collected from six cameras, in which each identity is observed by two cameras and has 4.8 bounding boxes on average in each camera. There are two bounding box generation versions which are manually labeled and automatically detected by the pedestrian detector DPM [16], respectively. In our experiment, we evaluate the automatically "detected" version. Following the protocol in [29], 1,360 identities are split into 1,160 identities for training, 100 identities for validation and 100 identities for testing. The initial learning rate is set to 0.001 and reduced by a factor of 0.1 after each 5 epochs. Training is done after 25 epochs. We report the averaged result after training/testing 20 times (following [29]) and use the single-shot setting. The testing process is performed in the cross-camera mode.

We evaluate our method based on two identification CNN models (i.e., CaffeNet [25] and ResNet-50 [19]), with a comparison with the baseline on CUHK03 dataset in Table 3. We observe from Table 3 that the rank-1 accuracy increases from 53.90 to 54.40% (+0.5%),

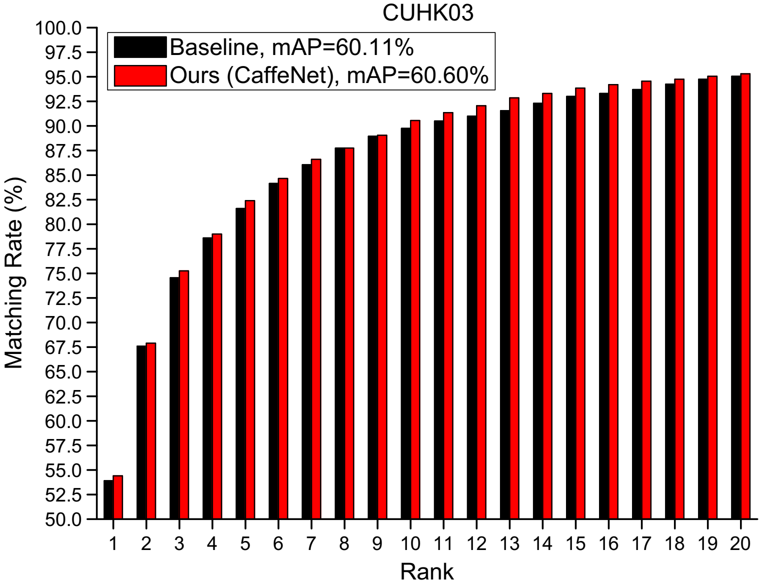

Fig. 5 The CMC of the baseline and our method comparison employing CaffeNet [25] model on CUHK03 dataset

the mAP increases from 60.11 to 60.60% (+0.49%) on CaffeNet model. On ResNet-50 model, both of the rank-1 accuracy and mAP have obviously improvement, the rank-1 accuracy increases from 54.50 to 56.65% (+2.15%), the mAP increases from 60.72 to 61.97% (+1.25%). Figures 5 and 6 summarize the CMC of the baseline and our method comparison on the two CNN models, respectively. The CMC has a stable improvement from rank-1 to rank-20 on CaffeNet and ResNet-50 models.

Comparison with the state-of-the-art re-ID methods We first compare our method with the conventional non-deep learning method. The features include the BoW [63] descriptor and LOMO [30]. The metric learning methods include LMNN [47], ITML [12], KISSME [24] and XQDA [30]. As can be seen in Table 4, compared with the conventional non-deep learning method, our CNN model brings a significant improvement of benchmark in rank-1 accuracy on CUHK03 dataset. Then we compare with some state-of-the-art person re-ID methods based on deep learning, including FPNN [29], DML [58], Improved Siamese [1] and SI-CI [45]. From the results in Table 4, it is clear that our method outperforms most of deep learning method in rank-1 accuracy on CUHK03 dataset. But the advantage of identification CNN model is restricted by the limitation amount (9.6 bounding boxes on average) of training samples for each identity on CUHK03 dataset. The experimental results still demonstrate the effectiveness of our method once again that the softmax loss and center loss are jointly employed as supervision signal for person re-ID task.

4.3 Experiments on MARS dataset

The MARS [65] dataset is a video-based benchmark for person re-ID, which contains 20,175 tracklets of 1,261 identities. The tracklet is generated by the DPM detector [16] and

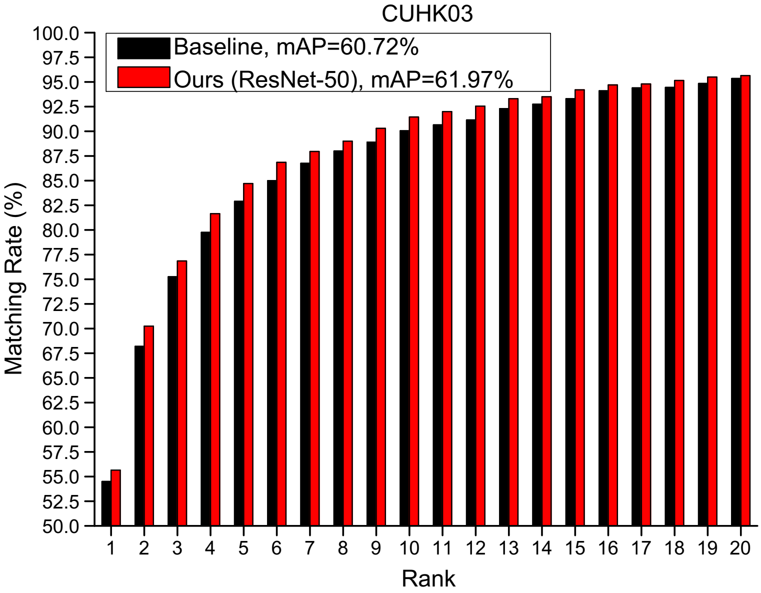

Fig. 6 The CMC of the baseline and our method comparison employing ResNet-50 [19] model on CUHK03 dataset

Table 4 Comparison with the state-of-the-art person re-ID methods (rank-1 accuracy (%) and mAP (%)) on CUHK03 dataset

Methods	CUHK03	
	rank-1	mAP
BoW+HS [63]	24.33	–
BoW+LMNN [47]	6.25	–
BoW+ITML [12]	5.14	–
BoW+KISSME [24]	11.70	–
LOMO+XQDA [30]	46.25	–
FPNN [29]	19.89	–
DML [58]	49.84	–
Improved Siamese [1]	44.96	–
SI-CI [45]	52.17	–
Ours (CaffeNet [25])	54.40	60.60
Ours (ResNet-50 [19])	**55.65**	**61.97**

The experimental results of our method are highlighted in bold

the GMMCP tracker [13]. Compared to image-based person re-ID datasets, the amount of training data is clearly larger in MARS. There are 509,914 image sequences in the training set. The total 1,261 identities are split into 631 identities for training and 630 identities for testing, followed by the protocol in [65]. During the training process of CNN model, 10% of the samples in training set are randomly selected as validation set. Pedestrian image sequences are employed to train the CNN model and learn an ID-discriminative Embedding in person subspace for each identity. The initial learning rate is set to 0.001 and reduced by a factor of 0.1 after each 4 epochs. Training is done after 16 epochs. During testing, the deep features of image sequences are extracted by the trained CNN model. The final feature representation of each tracklet is pooled by corresponding deep features of image sequences. We implement average pooling and max pooling, respectively.

Since MARS is released recently, we have not an extensive comparison results with the state-of-the-art person re-ID methods. We evaluate our method based on the CaffeNet [25] model, with a comparison with the baseline on MARS in Table 5. We observe from Table 5 that the rank-1 accuracy increases from 55.20 to 59.80% (+4.60%), and an improvement can be seen from mAP, from 35.78 to 42.00% (+6.22%) using average pooling. With the max pooling, both of the rank-1 accuracy and mAP have also obviously improvement, the rank-1 accuracy increases from 53.69 to 58.69% (+5.00%), the mAP increases from 33.59 to 40.56% (+6.97%). The average pooling has an improvement of about 2% in both the rank-1 accuracy and mAP compared with max pooling. Figures 7 and 8 summarize the CMC comparison of the baseline and our method using average pooling and max pooling,

Table 5 Comparison of the baseline and our method (rank-1 accuracy (%) and mAP (%)) employing CaffeNet [25] model on MARS datasets

Methods	Average pooling		Max pooling	
	rank-1	mAP	rank-1	mAP
Baseline	55.20	35.78	53.69	33.59
Ours	**59.80**	**42.00**	**58.69**	**40.56**

The experimental results of our method are highlighted in bold

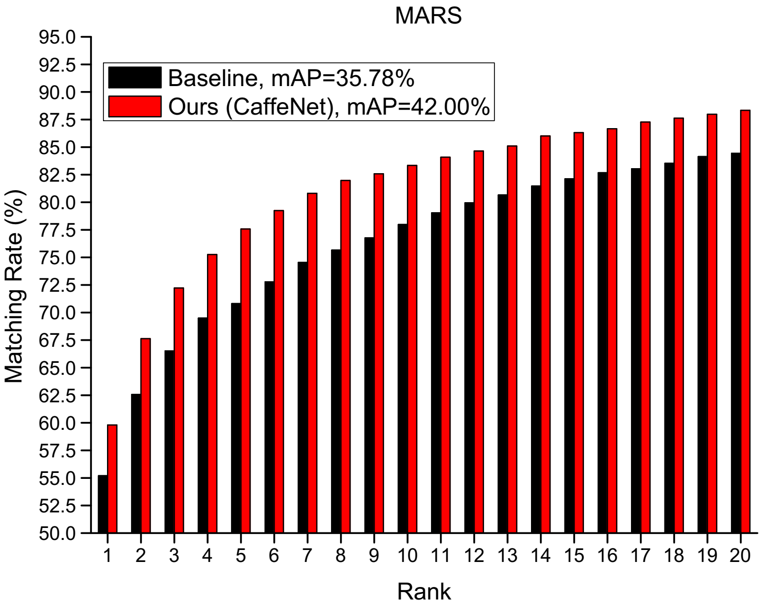

Fig. 7 The CMC of the baseline and our method comparison using average pooling on MARS dataset

respectively. The CMC has a significant improvement and the experimental results demonstrate the effectiveness of our method in video-based person re-ID task. Consequently, we believe that the research of video-based person re-ID still has potential improvement.

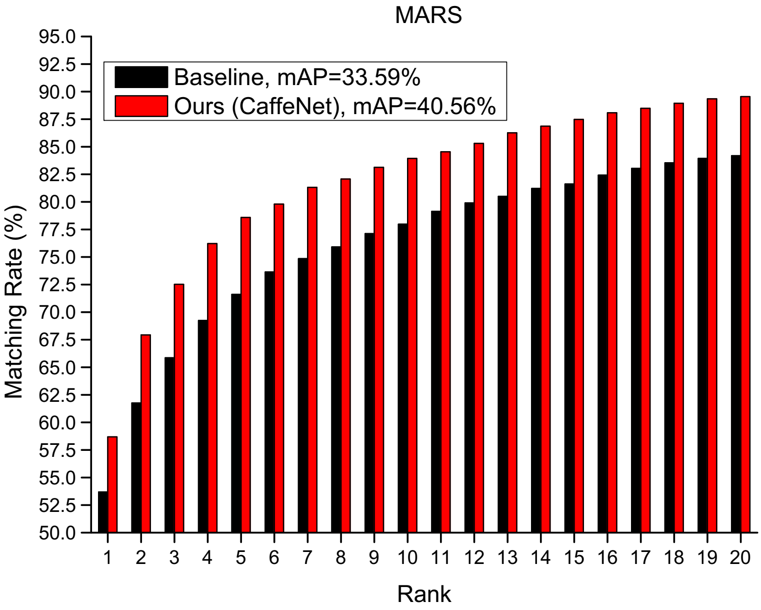

Fig. 8 The CMC of the baseline and our method comparison using max pooling on MARS dataset

5 Conclusion

In this paper, we combine the center loss with the softmax loss jointly to supervise the training process of the identification CNN model for person re-ID task. The softmax loss and center loss are combined to encourage the learned deep features of inter-class dispersed separably and intra-class aggregated compactly. The discriminative ability of the trained CNN model can be significantly improved. Extensive experiments on three important person re-ID benchmarks (i.e., Market-1501, CUHK03 and MARS) have convincingly demonstrated the effectiveness of our method. It yields a competitive performance compared with state-of-the-art person re-ID methods.

There are several challenging directions along which we will extend this work. First, larger databases with millions of pedestrian images that contain greatly intra-class and inter-class variances will be built to fully show the strength of deep learning methods. Second, more discriminative CNN models will be investigated to learn effective pedestrian representations.

Acknowledgements This work is supported by the Foundation for Innovative Research Groups of the NSFC (Grant no.71421001), National Natural Science Foundation of China (Grant no.61502073), and the Open Projects Program of National Laboratory of Pattern Recognition (No.201407349).

References

1. Ahmed E, Jones M, Marks TK (2015) An improved deep learning architecture for person re-identification. In: Proceedings CVPR, pp 3908–3916
2. An L, Chen X, Liu S, Lei Y, Yang S (2016) Integrating appearance features and soft biometrics for person re-identification. Multimedia Tools and Applications
3. Baltieri D, Vezzani R, Cucchiara R (2011) 3dpes: 3d people dataset for surveillance and forensics. In: Proceedings ACM workshop on human gesture and behavior understanding, pp 59–64
4. Bottou L (2010) Large-scale machine learning with stochastic gradient descent. In: Proceedings COMPSTAT'2010, pp 177–186
5. Chang X, Yang Y (2016) Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans Neural Netw Learn Syst. doi:10.1109/TNNLS.2016.2582746
6. Chang X, Nie F, Wang S, Yang Y, Zhou X, Zhang C (2016) Compound rank-*k* projections for bilinear analysis. IEEE Trans Neural Netw Learn Syst 27(7):1502–1513
7. Chang X, Nie F, Yang Y, Zhang C, Huang H (2016) Convex sparse pca for unsupervised feature learning. ACM Trans Knowl Discov Data 11(1):3:1–3:16
8. Chang X, Yu YL, Yang Y, Xing EP (2016) Semantic pooling for complex event analysis in untrimmed videos. IEEE Trans Pattern Anal Mach Intell. doi:10.1109/TPAMI.2016.2608901
9. Chen D, Yuan Z, Chen B, Zheng N (2016) Similarity learning with spatial constraints for person re-identification. In: Proceedings CVPR, pp 1268–1277
10. Cheng D, Gong Y, Zhou S, Wang J, Zheng N (2016) Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings CVPR, pp 1335–1344
11. Das A, Chakraborty A, Roy-Chowdhury AK (2014) Consistent re-identification in a camera network. In: Proceedings ECCV, pp 330–345
12. Davis JV, Kulis B, Jain P, Sra S, Dhillon IS (2007) Information-theoretic metric learning. In: Proceedings ICML, pp 209–216
13. Dehghan A, Modiri AssariS, Shah M (2015) Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: Proceedings CVPR, pp 4091–4099
14. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings CVPR, pp 248–255
15. Ess A, Leibe B, Van Gool L (2007) Depth and appearance for mobile scene analysis. In: Proceedings ICCV, pp 1–8

16. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D (2010) Object detection with discriminatively trained part-based models. IEEE Trans Pattern Anal Mach Intell 32(9):1627–1645
17. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings CVPR, pp 580–587
18. Gray D, Tao H (2008) Viewpoint invariant pedestrian recognition with an ensemble of localized features. In: Proceedings ECCV, pp 262–275
19. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings CVPR, pp 770–778
20. Hirzer M, Beleznai C, Roth PM, Bischof H (2011) Person re-identification by descriptive and discriminative classification. In: Proceedings Scandinavian conference on Image analysis, pp 91–102
21. Hu HM, Fang W, Zeng G, Hu Z, Li B (2016) A person re-identification algorithm based on pyramid color topology feature. Multimedia Tools and Applications
22. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, Guadarrama S, Darrell T (2014) Caffe: Convolutional architecture for fast feature embedding. In: Proceedings ACM international conference on multimedia, pp 675–678
23. Jose C, Fleuret F (2016) Scalable metric learning via weighted approximate rank component analysis. In: Proceedings ECCV, pp 875–890
24. Koestinger M, Hirzer M, Wohlhart P, Roth PM, Bischof H (2012) Large scale metric learning from equivalence constraints. In: Proceedings CVPR, pp 2288–2295
25. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings NIPS, pp 1097–1105
26. Leng Q, Hu R, Liang C, Wang Y, Chen J (2015) Person re-identification with content and context re-ranking. Multimedia Tools Appl 74(17):6989–7014
27. Li W, Wang X (2013) Locally aligned feature transforms across views. In: Proceedings CVPR, pp 3594–3601
28. Li W, Zhao R, Wang X (2012) Human reidentification with transferred metric learning. In: Proceedings ACCV, pp 31–44
29. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings CVPR, pp 152–159
30. Liao S, Hu Y, Zhu X, Li SZ (2015) Person re-identification by local maximal occurrence representation and metric learning. In: Proceedings CVPR, pp 2197–2206
31. Lin Y, Zheng L, Zheng Z, Wu Y, Yang Y (2017) Improving person re-identification by attribute and identity learning. arXiv:170307220
32. Liu H, Feng J, Qi M, Jiang J, Yan S (2016) End-to-end comparative attention networks for person re-identification. arXiv:160604404
33. Liu J, Zha ZJ, Tian Q, Liu D, Yao T, Ling Q, Mei T (2016) Multi-scale triplet cnn for person re-identification. In: Proceedings ACM international conference on multimedia, pp 192–196
34. Martinel N, Das A, Micheloni C, Roy-Chowdhury AK (2016) Temporal model adaptation for person re-identification. arXiv:160707216
35. Prosser B, Zheng WS, Gong S, Xiang T, Mary Q (2010) Person re-identification by support vector ranking. In: Proceedings BMVC, pp 1–11
36. Radenović F, Tolias G, Chum O (2016) Cnn image retrieval learns from bow: unsupervised fine-tuning with hard examples. arXiv:160402426
37. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv:151106434
38. Roth PM, Hirzer M, Koestinger M, Beleznai C, Bischof H (2014) Mahalanobis distance learning for person re-identification. In: Person re-identification, Springer London, pp 247–267
39. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: Proceedings CVPR, pp 815–823
40. Su C, Zhang S, Xing J, Gao W, Tian Q (2016) Deep attributes driven multi-camera person re-identification. In: Proceedings ECCV, pp 475–491
41. Sun Y, Zheng L, Deng W, Wang S (2017) Svdnet for pedestrian retrieval. arXiv:170305693
42. Ustinova E, Ganin Y, Lempitsky V (2015) Multiregion bilinear convolutional neural networks for person re-identification. arXiv:151205300
43. Varior RR, Haloi M, Wang G (2016) Gated siamese convolutional neural network architecture for human re-identification. In: Proceedings ECCV, pp 791–808
44. Varior RR, Shuai B, Lu J, Xu D, Wang G (2016) A siamese long short-term memory architecture for human re-identification. In: Proceedings ECCV, pp 135–153

45. Wang F, Zuo W, Lin L, Zhang D, Zhang L (2016) Joint learning of single-image and cross-image representations for person re-identification. In: Proceedings CVPR, pp 1288–1296
46. Wang T, Gong S, Zhu X, Wang S (2014) Person re-identification by video ranking. In: Proceedings ECCV, pp 688–703
47. Weinberger KQ, Blitzer J, Saul LK (2005) Distance metric learning for large margin nearest neighbor classification. In: Proceedings NIPS, pp 1473–1480
48. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: Proceedings ECCV, pp 499–515
49. Wu L, Shen C, Hengel AvD (2016) Personnet: Person re-identification with deep convolutional neural networks. arXiv:160107255
50. Xiang ZJ, Chen Q, Liu Y (2014) Person re-identification by fuzzy space color histogram. Multimedia Tools Appl 73(1):91–107
51. Xiao T, Li H, Ouyang W, Wang X (2016) Learning deep feature representations with domain guided dropout for person re-identification. In: Proceedings CVPR, pp 1249–1258
52. Xiao T, Li S, Wang B, Lin L, Wang X (2016) End-to-end deep learning for person search. arXiv:160401850
53. Yan Y, Ni B, Song Z, Ma C, Yan Y, Yang X (2016) Person re-identification via recurrent feature aggregation. In: Proceedings ECCV, pp 701–716
54. Yan Y, Nie F, Li W, Gao C, Yang Y, Xu D (2016) Image classification by cross-media active learning with privileged information. IEEE Trans Multimedia 18(12):2494–2502
55. Yang Y, Zhuang YT, Wu F, Pan YH (2008) Harmonizing hierarchical manifolds for multimedia document semantics understanding and cross-media retrieval. IEEE Trans Multimedia 10(3):437–446
56. Yang Y, Nie F, Xu D, Luo J, Zhuang Y, Pan Y (2012) A multimedia retrieval framework based on semi-supervised ranking and relevance feedback. IEEE Trans Pattern Anal Mach Intell 34(4):723–742
57. Yang Y, Ma Z, Hauptmann AG, Sebe N (2013) Feature selection for multimedia analysis by sharing information among multiple tasks. IEEE Trans Multimedia 15(3):661–669
58. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: Proceedings ICPR, pp 34–39
59. Zhang L, Xiang T, Gong S (2016) Learning a discriminative null space for person re-identification. In: Proceedings CVPR, pp 1239–1248
60. Zhao R, Ouyang W, Wang X (2014) Learning mid-level filters for person re-identification. In: Proceedings CVPR, pp 144–151
61. Zhao Y, Zhao X, Luo R, Liu Y (2016) Person re-identification by encoding free energy feature maps. Multimedia Tools Appl 75(8):4795–4813
62. Zheng L, Wang S, Liu Z, Tian Q (2014) Packing and padding: coupled multi-index for accurate image retrieval. In: Proceedings CVPR, pp 1939–1946
63. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings ICCV, pp 1116–1124
64. Zheng L, Wang S, Tian L, He F, Liu Z, Tian Q (2015) Query-adaptive late fusion for image search and person re-identification. In: Proceedings CVPR, pp 1741–1750
65. Zheng L, Bie Z, Sun Y, Wang J, Wang S, Su C, Tian Q (2016) Mars: a video benchmark for large-scale person re-identification. In: Proceedings ECCV, pp 868–884
66. Zheng L, Wang S, Wang J, Tian Q (2016) Accurate image search with multi-scale contextual evidences. Int J Comput Vis 120(1):1–13
67. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification: past, present and future. arXiv:161002984
68. Zheng L, Yang Y, Tian Q (2017) Sift meets cnn: a decade survey of instance retrieval. IEEE Trans Pattern Anal Mach Intell. doi:10.1109/TPAMI.2017.2709749
69. Zheng L, Zhang H, Sun S, Chandraker M, Yang Y, Tian Q (2017) Person re-identification in the wild. In: Proceedings CVPR
70. Zheng WS, Gong S, Xiang T (2009) Associating groups of people. In: Proceedings BMVC, pp 23.1–23.11
71. Zheng Z, Zheng L, Yang Y (2017) Unlabeled samples generated by gan improve the person re-identification baseline in vitro. arXiv:170107717

Fuqing Zhu received his B.E. and M.S. degrees from Dalian Jiaotong University, China, in 2010 and 2013, respectively. Currently, he is seeking his Ph.D. degree in School of Information and Communication Engineering at Dalian University of Technology, China. His research interests include multimedia retrieval, image classification and person re-identification.

Xiangwei Kong received her Ph.D. degree in Management Science and Engineering from Dalian University of Technology, China, in 2003. From 2006 to 2007, she was a visiting researcher in Department of Computer Science at Purdue University, USA. She is currently a professor in the School of Information and Communication Engineering at Dalian University of Technology, China. Her research interests include digital image processing and recognition, multimedia information security, digital media forensics, image retrieval and mining, multisource information fusion, knowledge management and business intelligence.

Qun Wu is an Assistant Professor in Product Innovation Design, Zhejiang Sci-Tech University, China. He received his Ph.D. from the College of Computer Science and Technology from Zhejiang University in China. He holds a Bachelor degree in Industrial Design from Nanchang University in China, and a Masters degree in Mechanical Engineering from Shaanxi University of Science and Technology in China. His research interests include machine learning, human factor and product innovation design.

Haiyan Fu received her Ph.D. from Dalian University of Technology, China, in 2014. She is currently an associate professor in the School of Information and Communication Engineering at Dalian University of Technology, China. Her research interests are in the areas of image retrieval and computer vision.

Ming Li received the M.S. and Ph.D. degrees in electrical engineering from the State University of New York at Buffalo, Buffalo, in 2005 and 2010, respectively. From Jan. 2011 to Aug. 2013, he was a Post-Doctoral Research Associate with the Signals, Communications, and Networking Research Group, Department of Electrical Engineering, State University of New York at Buffalo. From Aug. 2013 to June 2014, Dr. Li joined Qualcomm Technologies Inc. as a Senior Engineer. Since June 2014, he has been with the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China, where he is presently an Associate Professor. His research interests are in the general areas of communication theory and signal processing with applications to interference channels and signal waveform design, secure wireless communications, cognitive radios and networks, data hiding and steganography, and compressed sensing.