J. Vis. Commun. Image R. 62 (2019) 267–278

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

Discriminative representation learning for person re-identification via multi-loss training ☆

Weilin Zhong [a], Tao Zhang [a], Linfeng Jiang [a], Jinsheng Ji [a], Zenghui Zhang [b,c], Huilin Xiong [a,b,c,*]

[a] School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, China
[b] Institute for Sensing and Navigation, Shanghai Jiao Tong University, China
[c] Shanghai Key Laboratory of Intelligent Sensing and Recognition, Shanghai Jiao Tong University, China

ARTICLE INFO

Article history:
Received 24 January 2019
Revised 2 May 2019
Accepted 2 June 2019
Available online 4 June 2019

Keywords:
Person re-identification
Multi-loss training
Inter-center loss

ABSTRACT

The identification model that employs softmax loss to minimize person identity classification errors has gradually gained popularity in person re-identification community due to its easy implementations. However, the softmax loss only encourages the separation of different identities. The intra-class differences caused by large view variations such as spatial misalignment and human pose change are not considered in the model training process. In this paper, we present a hybrid deep model that combines multiple loss functions to handle this problem. Specifically, the multi-loss function contains three terms, namely softmax loss, center loss, and a novel loss called inter-center loss. The center loss penalizes the distance between deep features and their center, aiming to reduce intra-class differences. The inter-center loss maximizes the distances between different class centers, aiming to further enlarge inter-class separation. Extensive experiments conducted on three public benchmark datasets including Market1501, CUHK03, and DukeMTMC-reID demonstrate the effectiveness of our method.

1. Introduction

Person re-identification (re-id) is the task of matching images of the same individual across non-overlapping cameras. It has attracted increasing interests in computer vision community for its wide applications in human retrieval, cross-camera tracking and so on. Person re-id is a challenging task because of two major aspects, namely large intra-class variances and small inter-class differences. First, as shown in Fig. 1(a), the same person across camera views undergo large appearance changes due to the view variations in camera viewpoints, human poses, background clutters, and illumination. Second, different persons have similar appearances in clothes and human poses as illustrated in Fig. 1 (b). Therefore, the inter-class differences of different persons may be much smaller than the intra-class variances of the same person. Most traditional methods address this problem by first extracting hand-crafted features to represent person images and then learning distance metrics for similarity calculation. Those methods separately optimize the process of feature extraction and metric learning, which may result in a suboptimal performance. Recently, the Convolutional Neural Networks (CNN) based methods have dominated the person re-id community and obtained large performance improvement. Different from traditional methods, deep CNN-based approaches essentially integrate feature extraction and metric learning into one unified framework.

There exist two major types of CNN models for person re-id, namely verification models and identification models. The two models are different in terms of data organization, feature extraction, and loss functions. In the first category, verification models usually take image triplets or pairs as input and use a Siamese CNN for feature extraction. Various loss functions including triplet loss [1,2], contrastive loss [3–5], and their variants [6,7] are proposed to learn a deep embedding, where similar examples are mapped close to each other, while dissimilar examples are pushed far apart. Those methods only require weak labels about whether two images coming from the same person and thus they do not take full use of annotation information. Besides, the number of image triplets or pairs grows cubically with the scale growth of datasets, which may result in slow convergence and poor local optima [7]. In the second category, identification models regard person re-id as a multi-class identity classification problem, which can fully exploit identity labels. But, identification models require a rich amount of samples per identity to overcome over-fitting. Recently, several large-scale datasets such as Market1501 [8] and DukeMTMC-reID [9] are collected, and some useful training tricks

☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author at: School of Electronic, Information and Electrical Engineering, Shanghai Jiao Tong University, China.
E-mail address: hlxiong@sjtu.edu.cn (H. Xiong).

https://doi.org/10.1016/j.jvcir.2019.06.001
1047-3203/

(a) Images in the dotted box belong to the same person identity

(b) Images in the dotted box belong to different person identities

Fig. 1. The challenges in person re-id. The intra-class differences are very large due to large view variations, such as human pose change, illumination variances, and occlusion. The inter-class variances are very small because of the similar appearances in clothing and human pose.

[9,10] are also proposed to build an effective identification CNN. On large-scale datasets, many studies [11–14] show that identification models without any special data sampling process can obtain rather high re-id accuracy. Therefore, we build our method upon identification model.

However, existing identification models mainly adopt softmax loss for model training, which only encourage the separation of different identities. As illustrated in Fig. 1(b), the large intra-class differences caused by view variations are not handled in the model training process. In the task of face recognition, a new loss function called center loss [15] is proposed to explicitly eliminate the intra-class differences. Center loss learns a parametric center for deep features of each class and simultaneously penalizes the distances between deep features and their corresponding center. The parametric center loss in [15] has two major drawbacks. First, it can not be independently optimized, otherwise the learnt centers and deep features will easily degrade to zeros. Second, the parametric center is not the real centroid of each class and it requires additional optimization algorithm to update center parameters [15], which is not efficient. Therefore, a non-parametric center loss is proposed in this work. Our center loss first computes the center of each class, and then, directly minimizes the distances between features and their centers. Besides, both softmax loss and center loss do not explicitly constrain the distances between different class centers. The discriminative power of deep features may be affected if different class centers get too close in the embedding space. The two losses have their respective advantages and limitations. Our motivation is to absorb the advantages of two losses, and meanwhile, add a new loss term to overcome their weaknesses.

In this paper, we combine multiple loss functions in an identification model to learn a joint deep embedding for person re-id task. The multi-loss function consists of three terms, namely softmax loss, center loss, and a novel inter-center loss. Specifically, softmax loss learns to separate different identities using a softmax classifier. In the meanwhile, center loss and inter-class are adopted to regularize the deep features. Center loss pushes the features of the same person towards their identity center, aiming to reduce intra-class differences. Inter-center loss is introduced to maximize the distances between different identity centers, aiming to further enlarge inter-class separation. Fig. 2 illustrates the effects of different loss functions. The contributions of this paper are summarized as below:

(1) We extend the parametric center loss in [15] into its non-parametric version by directly calculating the centroid of each person identity.
(2) A novel loss function that maximizes the distances between different class centers is presented to further enhance the feature separation ability of identification model.
(3) Multiple loss functions are combined in a hybrid deep model to learn more robust feature representations. Extensive experiments are conducted on three benchmark datasets to validate the effectiveness of loss combinations.

The rest of this paper are organized as follows: The related studies are reviewed in Section 2. Our method is described in Section 3 and the experimental results are presented in Section 4. Finally, we conclude this paper in Section 5.

2. Related work

Person re-id has received increasing attention due to its wide applications in video surveillance. Typically, a person re-id system contains two major components: (1) extracting robust features to represent the query and gallery person images, (2) learning distance metrics to measure the similarity between the extracted

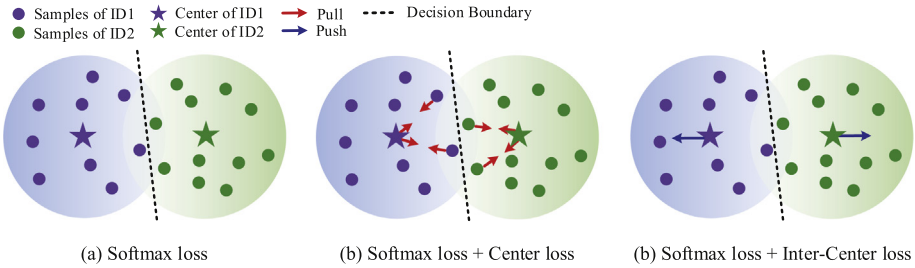


Fig. 2. A toy illustration of different loss functions and their combinations. Intuitively, the decision boundary of softmax classifier separates two classes. The center loss pulls features toward their corresponding identity center. The inter-center loss pushes away different identity centers. ID is the abbreviation of identity.

features. For feature representations, many methods fully exploit basic image descriptors, such as RGB, LAB, color names [16], the local binary patterns (LBP) [17–20], Gabor filter feature [20], color histogram and its variants [18,19,21] etc. For example, Liao et al. [22] constructed LOMO features via maximizing the horizontal occurrence of local descriptors. Li et al. [20] combined LBP, HSV color histogram, Gabor and HoG to represent person images. Gray et al. [23] utilized AdaBoost algorithm to learn the most discriminative features. Farenzena et al. [24] fully utilized the symmetry and asymmetry property of body structures. Kviatkovsky et al. [25] proposed an illumination-invariant color descriptor based on the color intra-distribution signatures. Lu et al. [26,27] proposed to learn local binary features for face recognition. For metric learning, many studies aim to find a mapping function from the feature space to distance space, where intra-class distances are minimized and inter-class distances are maximized. For instance, Davis et al. [28] proposed information-theoretic metric learning (ITML) method based on Mahalanobis distance. Zheng et al. [29] proposed a relative distance comparison (RDC) learning method from a probabilistic prospective. Liao et al. [22] proposed the Cross-view Quadratic Discriminant Analysis (XQDA), which simultaneously learnt a discriminant low-dimensional subspace and a distance metric. Other representative metric learning based methods include large scale metric learning from equivalence constraint (KISSME) [19], Local Fisher Discriminant Analysis (LFDA) [30], Large Margin Nearest Neighbor (LMNN) [31] and etc. Those methods mentioned above regard feature extraction and metric learning as two separated stages, whose performances may be limited.

Recently, the deep learning based methods that incorporate feature extraction and metric learning into one integrated framework have obtained large performance improvement in person re-id task. Many methods are based on verification models, which take triplets or paired images as input and use a distance layer outputting pairwise similarity. For example, Chen et al. [32] utilized a deep ranking framework to learn the joint representation of the horizontally stitched image pairs. Yi et al. [3] proposed a Siamese convolutional neural network followed by a cosine layer for similarity calculation. Ahmed et al. [33] proposed an improved deep learning architecture (IDLA) to compute the cross-input neighborhood differences of two images based on the mid-level features. Later, Wu et al. [34] improved the model in [33]. They increased the network depth using very small convolution filters. Varior et al. [4] utilized a long short-term memory (LSTM) architecture to model the spatial contextual information between different parts. Varior et al. [35] also proposed to capture the subtle difference between two images using a gated CNN. Cheng et al. [6] jointly learnt global full-body and local body-part features in a triplet framework. Lu et al. [36–38] proposed to learn a latent feature space, where the distance of each positive pair is reduced and that of each negative pair is enlarged. Duan et al. [39] proposed a deep adversarial metric learning (DAML) framework to generate synthetic hard negatives. Wang et al. [5] proposed to jointly learn single-image and cross-image representations in a unified framework. Despite yielding promising performances, verification models usually suffer from slow convergence due to large scale of image triplets or pairs. Moreover, the performances of verification models highly rely on hard negative [2,6] or hard positive [1,40] sample mining, which is also a time-consuming process.

Many other methods are based on identification models. The scale growth of person re-id datasets such as Market1501 and DukeMTMC-reID make it possible to train an identification model without over-fitting. On Market1501 dataset, there are on average 17.2 training images for each person identity. Identification models can learn good feature embeddings without complex data sampling process. For instance, Xiao et al. [41] combined multiple person re-id datasets to train an identification model and it performed many recent results. Zheng et al. [9] proposed to generate new training samples using a Generative Adversarial Network (GAN). Zhong et al. [12,42] proposed a novel camera style transfer model based on CycleGAN [43]. Sun et al. [10] utilized singular vector decomposition (SVD) to decorrelate the learnt weight vectors of identification CNN. Some methods use auxiliary information such as person attributes and human poses to assist identity classification. For example, Lin et al. [44] simultaneously predicted person identity and attributes in an attribute-person recognition (APR) network. Zhao et al. [45] utilized the Convolutional Pose Machines (CPM) [46] to localize head-shoulder region, upper body region, and lower body region. Zheng et al. [47] utilized CPM [46] to align the person images with spatial displacement to a standard pose. One major weakness of identification model lies in that the large intra-class differences are not considered in the model training stage. Hence, in this paper, we present a multi-loss function consisting of softmax loss, center loss, and inter-center loss to learn more discriminative features. The center loss aims to enhance intra-class compactness while the inter-center loss aims to further improve the feature separation capacity. Our model is also related to those methods using multi-loss functions for model training. For instance, Zheng et al. [13] jointly optimized verification loss and identification loss. Triplet loss and ranking loss are combined in [5]. Li et al. [48] proposed to take advantage of the complementary effects of global and local features in an identification model.

3. Proposed approach

In this work, we propose a hybrid deep model which solves person re-id as an identity classification problem. In this part, we will present our method in details. First, we illustrate the overall CNN architecture. Then, we introduce the three loss functions and combine them in the model training process. Finally, we discuss the training and testing strategies of our method.

3.1. The overall architecture

Fig. 3 illustrates the overall architecture of the proposed method. Our model is basically an identification network under the joint supervision of three loss functions. The loss combination

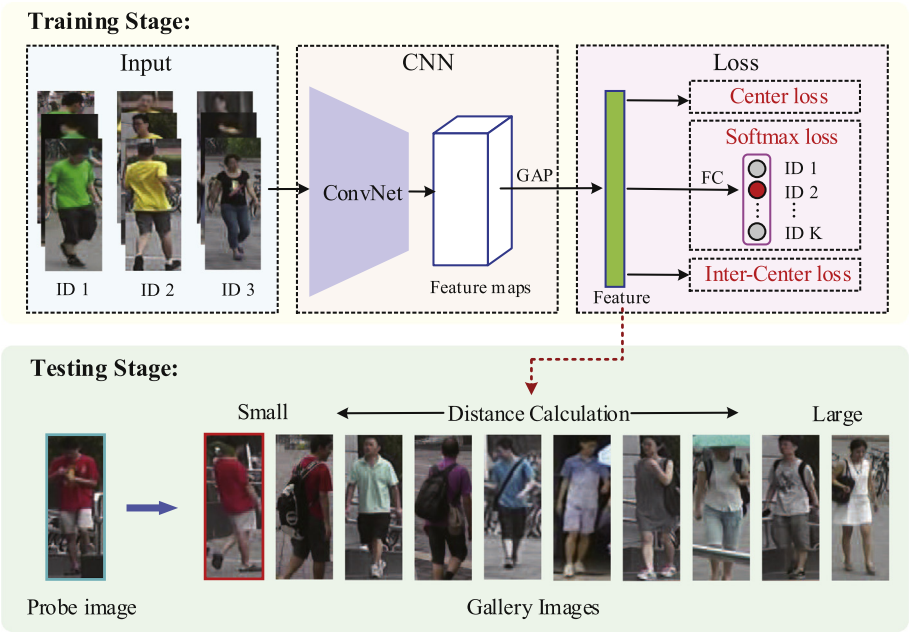


Fig. 3. Illustration of the proposed person re-id architecture. In the training stage, three losses functions are combined to guide the parameter updating process of CNN. In the testing stage, for each target person image, a ranking list is obtained according to their feature distances. GAP means global average pooling operation, which pools feature maps into feature vector. FC denotes fully connected layer, which acts as the identity classifier.

in the training process aims to learn discriminative features with inter-class separation and intra-class compactness.

The training stage contains three parts, namely the input part, the CNN part, and the loss part. First, the images from different person identities are sampled to construct a batch. Second, each image is forward into a backbone network to obtain its compact feature maps, which are then sent into a Global Average Pooling (GAP) layer to generate feature vector. The main purpose of our model is not to study the CNN structure but to learn a discriminative deep embedding using multiple functions. Thus, the deep CNN can be any off-the-shelf networks. Here, we use ResNet50 [49] as the backbone network due to its outstanding performances in ImageNet [50] classification. The structure of ResNet50 is shown in Table 1. Finally, in the loss part, three loss functions are combined to supervise the model training. Among them, softmax loss penalizes the identity classification errors, which learns an identity-separated embedding. In the meanwhile, center loss and inter-center loss regularize the deep features and identity centers so that more discriminative pedestrian features can be learnt.

3.2. Loss functions

As illustrated in Fig. 3, in the training stage, three loss functions including softmax loss, center loss, and inter-center loss are combined in our model to learn more stable pedestrian descriptors for person re-id task. Below we will show the details of different losses and present how to combine them.

Softmax loss. In a training batch, supposing that the number of images is N and each belongs to one of K person identities, softmax loss can be defined as:

$$\mathcal{L}_s = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{k=1}^{K} e^{W_k^T x_i + b_j}}, \tag{1}$$

Table 1
Network structure of ResNet50. There are five convolutional blocks in ResNet50. The first one is a convolutional layer and the rest four are residual blocks encapsulated with several convolutional layers. Here, ResNet50 takes 256×128 image as input and outputs 2048-dimension feature vector. We additionally add a batch normalization layer after the final GAP layer.

Layer name	Kernel size	Stride	Pad	Output size
conv1	$[7 \times 7, 64]$	2	3	$128 \times 64 \times 64$
Max pooling	$[3 \times 3]$	2	–	$64 \times 32 \times 64$
conv2	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times 3$	$64 \times 32 \times 256$
conv3	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	$\begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix} \times 4$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times 4$	$32 \times 16 \times 512$
conv4	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 6$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times 6$	$16 \times 8 \times 1024$
conv5	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2028 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \times 3$	$\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times 3$	$8 \times 4 \times 2048$
GAP	$[7 \times 7]$	1	–	2048
Batch normalization	–	–	–	2048

where $x_i \in R^d$ denotes the feature vector obtained from deep CNN and its corresponding ground-truth identity is y_i. d is the feature dimension. W_j represents the $j-th$ column of weights $W \in R^{d\times K}$ in the FC classifier and $b \in R^K$ is the bias term.

Center loss. Center loss pulls the deep features towards their corresponding identity centers. First, the center of each person identity is calculated:

$$c_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_j, \tag{2}$$

where n_k is the image number of $k-th$ identity in the training set.

Then, the distances between features and their identity centers are penalized, which can be expressed as:

$$\mathcal{L}_c = \frac{1}{N}\sum_{i=1}^{N}\|x_i - c_{y_i}\|_2^2 \tag{3}$$

where c_{y_i} represents the $y_i - th$ identity center.

Ideally, the center c_{y^i} should be constantly updated as the features change. In other words, according to Eq. (2), we need to take the entire training set into account and compute the centers of all identities in each iteration, which is inefficient and impractical. Therefore, similar to the parametric center loss in [15], here we simultaneously compute the center and minimize the distances between features and their corresponding identity centers in each training batch.

Inter-center loss. Inter-center loss maximizes the distances between different identity centers, which is formulated as:

$$\mathcal{L}_{ic} = \sum_{i\neq j}^{K}\frac{1}{\|c_i - c_j\|_2^2 + \delta} \tag{4}$$

where δ is a small constant used for numerical stability.

Similar to center loss, only the identity centers in the training batch are adopted for loss computation.

Loss combinations. The effects of different losses are shown in Fig. 2. Three loss functions have their respective strengthes and weaknesses. The softmax loss aims to find a decision boundary to separate different person identities. The deep features under the supervision of softmax loss may not be discriminative enough, because each identity still contain large intra-class differences. Therefore, it is not optimal to directly apply the deeply learnt features to person re-id task. Center loss can effectively enforce intra-class compactness via pulling features towards their identity centers. But, the distances between different identity centers are not constrained, which may deteriorate the re-id performances. Thus, inter-center loss is introduced to separate different identity centers. Similar to center loss, the main weakness of inter-center loss is that it can not be independently used. The three losses are complementary and can be combined to jointly optimize the identification model. Softmax loss enforces the inter-class separation, and meanwhile, center loss and inter-center loss regulates the deep features, so that more discriminative feature representations can be learnt.

The three losses are linearly combined, which can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_s + \lambda_c\mathcal{L}_c + \lambda_{ic}\mathcal{L}_{ic} \tag{5}$$

where λ_c and λ_{ic} are hyper-parameters which control the trade-off between different losses.

3.3. Training and testing

Training strategies. The images are first re-scaled into 288×140 and then they are cropped into 256×128. Data augmentation methods including mirror flip and minor rotation are applied to the cropped images. Besides, the pixels of all images are normalized to [0,1], subtracted by mean pixel values of RGB channels and then divided by standard deviation of each channel. Batch normalization and dropout operation are sequentially followed after the final GAP layer, which play a critical role in avoiding over-fitting. The dropout ratio is 50%.

We use PyTorch[1] to implement our method. In the training process, each batch consists of 16 person identities, and for each identity we randomly select 5 images, and thus the batch-size is 80. The model parameters are initialized from the ones pre-trained on ImageNet. Adam optimizer is applied to update the model parameters. The initial learning rate is set as 3×10^{-4}. The learning rate decreases every 50 epoches by a factor of 0.1 and the total training epoches is 180.

Testing strategies. On the testing phase, the test images are divided into two sets, namely probe and gallery sets. We feed forward the person images into the CNN architecture. The deep features of each pedestrian image is extracted from the last batch normalization layer. In this way, all the image features either from probe or gallery sets are obtained. Afterwards, Euclidean distance is adopted to calculate the distance between probe and gallery imaegs. For each probe image, the gallery images from the same person with same camera view are excluded. By sorting the distances between two sets, the final ranking list of cross-view images can be obtained.

4. Experiment results

4.1. Experimental settings

Datasets. We conduct the experiments on three public benchmark datasets, including Market1501 [8], CUHK03 [51], and DukeMTMC-reID [9]. Some examples from three datasets are displayed in Fig. 4. Market1501 is composed of 32,668 auto-detected bounding boxes of 1,501 identities. It is one of the largest datasets in the existing papers. Each person identity is captured by at most six cameras in front of a campus supermarket. The images are automatically detected by a pedestrian detector Deformable Part Model (DPM) [52]. CUKH03 contains 1,360 person identities and there are more than 13,000 images on this dataset. The images are collected from six cameras distributed at different locations in a university campus. Each identity is captured from two adjoint cameras and has 4.8 images on average for each view. Two versions are provided on this dataset, namely the manually labeled version and the automatically detected version by the DPM detector. We evaluate our model on the bounding boxes detected by DPM, which is closer to the realistic person re-id. DukeMTMC-reID dataset is a subset of the multi-target, multi-camera pedestrian tracking dataset [53]. We adopt the re-id version provided by [9], which consists of 1,401 person identities and 34,183 images. Each person is captured by at most eight cameras with non-overlapping views. The pedestrian bounding boxes are manually cropped.

Evaluation protocol. All the experiments are conducted under single query setting. Two widely used evaluation metrics are adopted for performance comparison, namely cumulative matching characteristic (CMC) [54] and mean average precision (mAP) [8]. Each dataset is split into two subsets, namely training and testing subset with non-overlapping person identities. For Market1501, we follow the standard training/testing protocol defined by [8], which uses fixed 750 identities as training subset and the rest fixed 751 identities as testing subset. For CUHK03, we adopt the new training/testing protocol proposed in [55], which fixes

[1] https://pytorch.org/.

(a) Market1501 (b) CUHK03 (c) DukeMTMC-reID

Fig. 4. Sample images from Market1501, CUHK03, and DukeMTMC-reID datasets.

767 person identities for training and the rest 700 identities for testing. For DukeMTMC, following the evaluation protocol in [9], 1,401 identities are divided into a training subset with 702 identities and a testing subset with the rest 702 identities.

4.2. Comparison with state-of-the-art methods

In this section, we compare the presented model with recent state-of-the-art methods, including conventional methods and deep learning based methods. Conventional methods usually are based on hand-crafted features. Those methods include SDALF [24], eSDC [56], BoW [8], KISSME [19] and LOMO [22]. The deep methods include PersonNet [34], End-to-end CAN [2], Siamese LSTM [4], ID-discriminative Embedding (IDE) [55,14], Gated CNN [35], Spindle Network (SpindleNet) [57], GAN [9], Pose Invariant Embedding (PIE) [47], CNN-Embedding [13], TriNet [40], Joint Learning Multi-Loss (JLML) [48], Pose-Sensitive Embedding (PSE) [58], Pedestrian Alignment Network (PAN) [59], Quadruplet [60], Cam-GAN [12,42], Harmonious Attention CNN (HA-CNN) [61], Dual Attention Matching network (DuATM) [62], SVDNet [10], Online Instance Matching (OIM) [63], and Attribute-Complementary Re-id Network (ACRN) [64]. Note that, not all of these approaches report their performances on all of the three datasets.

Results on Market1501. Market1501 is one of the largest benchmark dataset for person re-id and many methods have conducted experiments on this dataset. The experimental results are shown in Table 2. We can observe that the deep models, especially recent methods Cam-GAN [12,42], HA-CNN [61], and DuATM [62], outperform the hand-crafted features based methods (e.g., eSDC [56] and BoW [8]) by a large margin. This demonstrates the powerful feature learning capability of deep CNN. Compared to other deep models, our method obtains significantly better performance on this dataset. Specifically, our model achieves 92.15% Top1 accuracy and 79.73% mAP, which outperforms the previous best-performing method DuATM [62] by 0.73% (92.15% versus 91.42%) at Top1, and 3.22% (79.73% versus 76.62%) in mAP, respectively. Among the deep methods, PersonNet [34], CAN [2], Siamese LSTM [4], Gated CNN [35], Quadruplet [60], ACRN [64], HAP2S [65], and TriNet [40] are based on verification models. IDE [55,14], SpindleNet [57], GAN [9], PIE [47], JLML [48], PSE [58], Cam-GAN [12,42], HA-CNN [61], and DuATM [62] are based on identification models. Verification models either adopt complicated matching networks for part-level similarity calculation [34,4,35], or highly rely on mining of hard image pairs and triplets [2,60,65,40]. Compared to the verification models, the accuracies of our method are significantly better, with at least 7.23% Top1 accuracy improvement (ours 92.15% versus TriNet [40] 84.92%) and 10.59% mAP improvement (ours 79.73% versus TriNet [40] 69.14%). Besides, our method needs no special data selection process, which is more efficient. Compared to other identification models, our method requires no external resources, such as the human pose estimation model CPM [46] used in [57,47,58], camera information used in [12], attention mechanisms used in [61,62], and re-ranking scheme used in [58]. Moreover, our method performs better than those methods. For instances, our method outperforms PSE [58], Cam-GAN [12,42], and HA-CNN [61] by 4.45%, 4.03%, and 0.95% at Top1, and 10.73%, 11.01%, and 4.03% in mAP, respectively. In Fig. 5, we show the top 10 retrieved images from Market1501. It can be observed that our method exhibits strong robustness to scale variations, human pose change, and illumination variances. The failure cases are mainly caused by similar appearances in human pose and clothing. These false matchings are also very challenging from human perspective.

Table 2
Performance comparison on Market1501 dataset. "∗" denotes unpublished paper. "–" means no available reported results.

Methods	Top1	mAP
SDALF [24] (ICCV10)	20.5	8.2
eSDC [56] (CVPR13)	33.5	13.5
BoW [8] (ICCV16)	34.42	14.17
PersonNet*[34] (ArXiv16)	37.21	18.57
End-to-end CAN [2] (TIP17)	48.24	24.43
Siamese LSTM [4] (ECCV16)	61.6	35.31
IDE [55,14] (CVPR17)	72.54	46.03
Gated CNN [35] (ECCV16)	76.04	48.45
SpindleNet [57] (CVPR17)	76.9	–
GAN [9] (ICCV17)	78.06	56.23
PIE*[47] (ArXiv17)	78.65	53.87
CNN-Embedding [13]	79.51	59.87
Quadruplet [60] (CVPR17)	81.47	64.88
ACRN [64] (CVPRW17)	83.61	62.6
HAP2S [65] (ECCV18)	84.59	69.43
TriNet*[40] (ArXiv17)	84.92	69.14
JLML [48] (IJCAI17)	85.1	65.5
PSE [58] (CVPR18)	87.7	69.0
Cam-GAN [12,42] (CVPR18)	88.12	68.72
HA-CNN [61] (CVPR18)	91.2	75.7
DuATM [62] (CVPR18)	91.42	76.62
Our method	92.15	79.73

Results on CUHK03. CUHK03 dataset has two versions and we conduct experiments on the detected version, which is more closer the realistic re-id scenarios considering spatial displacement, and human pose changes. The experimental results are displayed in Table 3. It can be observed that our model achieves much better performance than all the hand-crafted features based methods, which include BoW [8] and LOMO [22]. On this dataset, our method obtains 59.29% Top1 accuracy and 57.84% mAP, which outperforms IDE(R) [55,14] and its combination with XQDA [22] by a

Fig. 5. Retrieval examples on Market1501 dataset using the proposed method. The images in the first column are the query images. The top 10 retrieved images are sorted according to the similarity scores from left to right. Red rectangles represent the correct matches.

Table 3
Performance comparison on CUHK03 detected dataset.

Methods	Top1	mAP
BoW [8] + KISSME [19]	6.4	–
LOMO [22] (CVPR15)	12.8	11.5
IDE(C) [55,14] (CVPR17)	12.8	–
IDE(C) [55,14] + XQDA [22]	21.1	–
IDE(R) [55,14] (CVPR17)	21.3	–
IDE(R) [55,14] + XQDA [22]	31.1	–
SVDNet [10] (ICCV17)	41.5	37.3
HA-CNN [61] (CVPR18)	41.7	38.6
PAN [59] (TCSVT18)	43.83	41.91
Our method	59.29	57.84

Table 4
Performance comparison on DukeMTMC-reID dataset.

Methods	Top1	mAP
BoW[8] + KISSME [19]	25.1	12.2
LOMO [22] (CVPR15)	30.83	17.24
GAN [9] (ICCV17)	67.68	47.13
OIM [63] (CVPR17)	68.1	–
ACRN [64] (CVPRW17)	72.58	51.96
Quadruplet [60] (CVPR17)	73.47	54.29
Cam-GAN [12,42] (CVPR18)	75.27	53.48
HAP2S [65] (ECCV18)	75.94	60.64
PAN [59] (TCSVT18)	75.94	66.74
SVDNet [10] (ICCV17)	76.7	56.8
PSE [58] (CVPR18)	79.8	62.0
DuATM [62] (CVPR18)	81.82	64.58
Our method	83.23	67.95

large margin. Compared to other deep methods, our model performs significantly better. For example, our method outperforms SVDNet [10], HA-CNN [61], and PAN [59] by 17.79%, 17.59%, and 15.46% at Top1, and 20.54%, 19.24%, and 15.93% in mAP, respectively. It is worth noticing that HA-CNN [61] combines many attention networks including spatial attention, channel attention, and hard regional attention for feature extraction. PAN [59] deploys Spatial Transformer Network (STN) [66] for person body alignment. Our method only extracts features from a single backbone network ResNet50. Fig. 6 shows some retrieval samples on CUHK03. It can be seen that all the images of the same persons from different cameras can be retrieved and the other retrieved images are also very similar to the target persons in appearances.

Results on DukeMTMC-reID. The comparison experiments are shown in Table 4. It can be seen that our model performs much better than most deep learning based methods. For example, our method improves the Top1 accuracy by 15.55%, 15.13%, 9.76%, and 7.29% over GAN [9], OIM [63], Quadruplet [60], and HAP2S [65], respectively. Our method obtains 83.23% Top1 accuracy and 67.95% mAP, which performs better than ACRN [64] that relies on person attributes, Cam-GAN [12,42] that requires camera information, and PSE [58] that utilizes auxiliary human pose cues and re-ranking scheme. Compared to PAN [59], our method improves the Top1 accuracy by 7.29%, and mAP by 1.21%. Besides, the accuracies of our model are better than SVDNet [10] and DuATM [62]. The respective Top1 accuracy improvements are 6.53% and 1.41%, and the respective mAP improvements are 11.15% and 3.37%. In

Fig. 6. Retrieval examples on CUHK03 detected dataset using the proposed method.

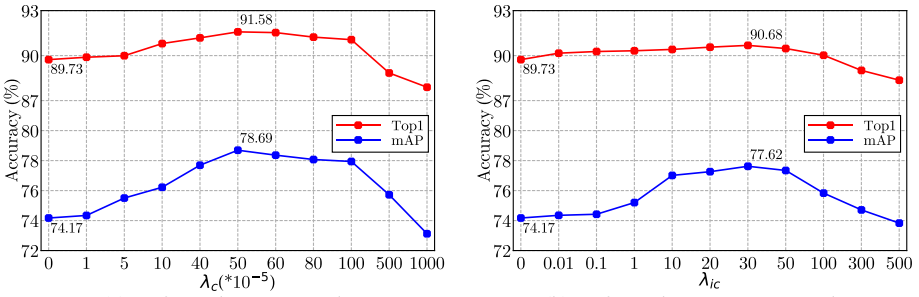
Fig. 7, we show some retrieval samples from DukeMTMC-reID dataset. In the third row, the wrongly matched person images are visually similar and there exist only subtle differences such as backpacks between them.

4.3. Model analysis

The roles of center loss and inter-center loss are different in our model. Center loss encourages intra-class compactness while inter-center loss enforces inter-class separation. Each of them can be applied with softmax loss to supervise model training. Therefore, we first study the effects of combing two losses: (1) $\mathcal{L}_s + \lambda_c \mathcal{L}_c$, (2) $\mathcal{L}_s + \lambda_{ic} \mathcal{L}_{ic}$. Then, we analyze the performance gains of combing three losses, namely $\mathcal{L}_s + \lambda_c \mathcal{L}_c + \lambda_{ic} \mathcal{L}_{ic}$.

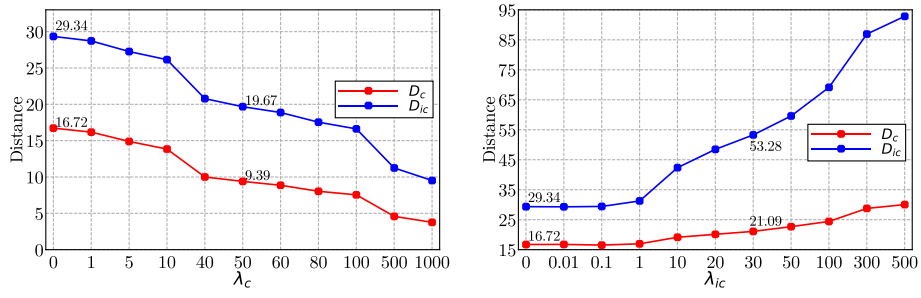
Combining two losses. In Figs. 8 and 9, we show the experimental results of combining two losses under different trade-off parameters λ_c and λ_{ic}. Fig. 8 displays the Top1 accuracy and mAP. Fig. 9 shows the averaged Euclidean distance between features and their corresponding center, which is $D_c = \frac{1}{N}\sum_{i=1}^{N}\|x_i - c_{y_i}\|_2$, and the averaged Euclidean distance between different class centers, which is $D_{ic} = \frac{1}{C_K^2}\sum_{i\neq j}^{K}\|c_i - c_j\|_2$. From Fig. 8(a), it can be observed that, as λ_c increases, the Top1 accuracy and mAP of combining softmax loss and center loss increase simultaneously until they reach their maximums at $\lambda_c = 5 \times 10^{-4}$, which are 91.58% and 78.69%, respectively. If λ_c is too large, the Top1 accuracy and mAP will drop. Especially, when λ_c is larger than $\lambda_c = 5 \times 10^{-3}$, the Top1 accuracy of combining softmax loss and center loss is lower than that of individual softmax loss, namely $\lambda_c = 0$. In Fig. 8(b), as λ_{ic} increases, the accuracy change tendency of combining softmax loss and inter-center loss is similar to that of combing softmax loss and center loss. At $\lambda_{ic} = 30$, the best Top1 accuracy and mAP are obtained, which are 90.68% and 76.62%, respectively. Importantly, from Fig. 9a) and (b), it can be seen that center loss can effectively reduce the distance between features and their corresponding center D_c, and inter-center loss can effectively enlarge

Fig. 7. Retrieval examples on DukeMTMC-reID dataset using the proposed method.



(a) Softmax loss+Center loss (b) Softmax loss+Inter-Center loss

Fig. 8. Top1 accuracy and mAP of combining two losses on Market1501 dataset.



(a) Softmax loss+Center loss (b) Softmax loss+Inter-Center loss

Fig. 9. The changes of the averaged Euclidean distance between features and their corresponding class center D_c, and the averaged Euclidean distance between different class centers D_{ic} on Market1501 dataset when combining two losses.

the distance between different class centers D_{ic}. But at the same time, center loss also reduces D_{ic}, and inter-center loss also enlarges D_c, which is undesirable in person re-identification. The two losses are thus simultaneously used to regulate the features learnt by softmax loss.

Combining three losses. In Table 5, we conduct experiments to determine the values of λ_c and λ_{ic} in three loss combinations. We observe that the best trade-off parameters $\lambda_c = 5^{-4}$ and $\lambda_{ic} = 30$ in two loss combinations obtain rather low accuracy in three loss combinations, because the regularization effects of two losses are different. Center loss reduces D_c but also reduces D_{ic}. Inter-center loss enlarges D_{ic} but also enlarges D_c. The two losses interact in the model training process and they are sensitive to the trade-off parameters. Even though, it can be observed that three loss combinations generally outperform two loss combinations if the trade-off parameters are properly tuned. This shows that softmax loss, center loss, and inter-center loss are complementary in nature. The best performances of three loss combination are obtained at $\lambda_c = 6 \times 10^{-4}$ and $\lambda_{ic} = 0.01$, which are 92.15% Top1 accuracy and 79.73% mAP. We thus use this setting to conduct experiments on other two datasets including CUHK03 and DukeMTMCre-ID. The experimental results are shown in Table 6. We can see that the loss combinations consistently improve the re-id performances on the two datasets. The most significant accuracy improvement comes from CUHK03 dataset, where softmax baseline only achieves 47.64% Top1 accuracy and 45.23% mAP. The loss combinations obtain at least 9.29% improvement in terms of Top1 accuracy (softmax + center 56.93% versus softmax 47.64%), and 10.76% improvement in terms of mAP (softmax + center 55.99% versus softmax 45.23%).

Moreover, in Table 6, we also show the performances of the parametric versions of center loss and inter-center loss. Their formulas are same to our non-parametric ones in Eq. (4) and (5), except that the centers are learnable parameters. Here, adam optimizer is used to update center parameters and the learning rate is set as 5×10^{-5}. We denote softmax loss, parametric center loss, and parametric inter-center loss as "S", "PC", and "PIC", respectively. In "S"+"PIC", the "PIC" has no regularization effects on features and thus it is equivalent to using "S" individually. The trade-off parameter for "S"+"PC" is tuned as $\lambda_{pc} = 4 \times 10^{-4}$, while for "S"+"PC"+"PIC", the best trade-off parameters are $\lambda_{pc} = 5 \times 10^{-4}$ and $\lambda_{pic} = 0.02$. From Table 6, it can be seen that the loss combinations, except "S"+"PIC", perform better than softmax baseline on three datasets. Three loss combination "S"+"PC"+"PIC" generally outperforms two loss combination "S"+"PC", which further validates the complementary attributes of center loss and inter-center loss. We also observe that our loss combinations, including "S"+"C", "S"+"IC", and "S"+"C"+"IC", obtain better accuracy than their parametric ones. Besides, instead of learning parametric centers, our method directly computes class centers, which is more efficient.

Experimental results on different CNN models. In Table 7, we conduct experiments on two extra widely used CNN models, namely AlexNet [50] and DenseNet121 [67]. Similar to ResNet50, all the FC layers of the two backbone networks are removed and the size of input image are same, which are 256×128. The output

Table 5
Experimental results of combing softmax loss, center loss ($\lambda_c * 10^{-5}$), and inter-center loss (λ_{ic}) on Market1501 dataset.

λ_c	λ_{ic}	Top1	mAP	D_c	D_{ic}	λ_c	λ_{ic}	Top1	mAP	D_c	D_{ic}
0	0	89.73	74.17	16.72	29.34	50	0	91.58	78.69	9.36	19.67
0	0.01	90.16	74.35	16.75	29.29	50	0.01	**91.87**	**79.56**	9.39	19.75
0	0.1	90.27	74.42	16.52	29.41	50	0.1	91.25	78.84	9.44	19.88
0	1	90.32	75.20	16.95	31.25	50	1	90.83	77.91	9.69	21.67
0	30	**90.68**	**77.62**	21.09	53.28	50	30	89.10	76.34	13.48	39.84
0	50	90.47	77.34	22.67	59.59	50	50	88.90	75.73	14.82	45.46
5	0	89.99	75.50	14.89	27.26	60	0	91.53	78.36	8.76	18.87
5	0.01	**91.63**	78.23	15.09	27.63	60	0.01	**92.15**	**79.73**	8.86	18.94
5	0.1	91.24	**78.82**	15.15	27.83	60	0.1	91.34	78.85	9.01	19.27
5	1	90.91	78.64	15.28	29.57	60	1	90.59	78.12	9.28	20.76
5	30	90.98	77.72	19.54	51.02	60	30	90.71	77.29	13.21	38.91
5	50	90.51	77.62	20.76	57.18	60	50	88.75	75.84	14.25	43.91
10	0	90.86	76.22	13.84	26.14	100	0	91.06	77.94	7.54	16.62
10	0.01	91.66	77.30	13.73	26.06	100	0.01	**91.37**	**78.20**	7.50	16.54
10	0.1	**92.04**	**78.95**	13.71	26.19	100	0.1	91.18	77.89	7.47	16.75
10	1	91.15	78.43	14.05	28.06	100	1	89.88	77.69	8.03	18.67
10	30	90.62	77.69	18.03	48.93	100	30	88.18	75.02	11.72	35.31
10	50	89.67	77.24	19.39	54.95	100	50	87.86	73.29	12.72	40.10

Table 6
Performance comparison of different losses on several datasets. The CMC Top1 accuracy (%) and mAP (%) are presented. S, C, and IC denote softmax loss, center loss, and inter-center loss, respectively. PC and PIC represent the parametric versions of center loss and inter-center loss, namely parametric center loss and parametric inter-center loss, respectively

Different Losses	Market1501		CUHK03		DukeMTMC-reID	
	Top1	mAP	Top1	mAP	Top1	mAP
S	89.73	74.17	47.64	45.23	80.30	62.99
S + C	91.58	78.69	56.93	55.99	82.81	67.10
S + IC	90.68	77.62	57.36	56.72	82.09	67.06
S + C+IC	92.15	79.73	59.29	57.84	83.23	67.95
S + PC	91.24	77.56	55.86	52.92	81.57	65.28
S + PIC	89.73	74.17	47.64	45.23	80.30	62.99
S + PC + PIC	91.53	78.02	57.12	54.80	82.06	66.14

Table 7
The performances of different losses on different deep CNN models over Market1501 dataset.

Different Losses	AlexNet		ResNet50		DenseNet121	
	Top1	mAP	Top1	mAP	Top1	mAP
S	68.76	42.08	89.73	74.17	90.11	74.42
S + C	71.88	45.55	91.58	78.69	91.23	76.68
S + IC	72.30	46.23	90.68	77.62	91.10	76.24
S + C+IC	73.37	47.73	92.15	79.73	92.04	78.85

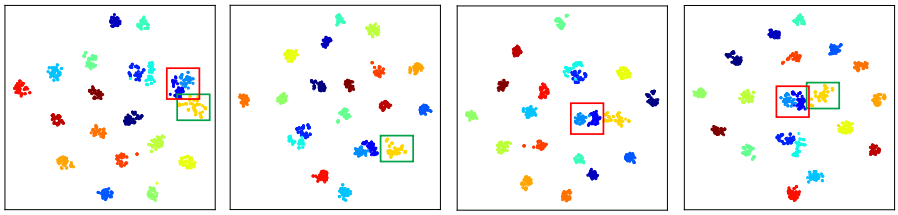


Fig. 10. The Barnes-Hut t-SNE visualization of the learnt features on Market1501 dataset. Here we randomly select 20 identities (each includes more than 40 images) from the gallery set for visualization. The same color represents the images coming from the same identity.

Table 8
Comparison of average feature extraction time and average retrieval time on Market1501 (millisecond per image).

Methods	Dimension	Feature extraction	Retrieval	Total
IDE(R) [55]	2048	5.8	6.33	12.13
CNN-Embedding [13]	2048	5.9	6.42	12.32
IDLA [33]	500	410.6	6.87	417.47
RCN [69]	4800	93.5	6.71	100.21
Our method	2048	6.1	6.48	12.58

feature maps of two CNN models are in size of $7 \times 3 \times 256$ and $8 \times 4 \times 1024$, respectively. Then, these feature maps are sequentially passed through GAP operation, batch normalization layer, and dropout layer, generating 256-dimension and 1024-dimension feature vectors, respectively. The training settings of DenseNet121 are same to ResNet50, while for AlexNet we use a slightly higher learning rate, which is 10^{-3}. From Table 7, it can be observed that the loss combinations consistently obtain performance gains on different backbone networks, which further demonstrates the effectiveness of the presented method in learning more robust features. In particular, the accuracy improvements on AlexNet are very significant. For example, compared to the softmax baseline, combining three losses achieves 4.61% Top1 accuracy improvement (73.77% versus 68.76%), and 5.65% mAP improvement (47.73% versus 42.08%).

Visualization. In Fig. 10, we adopt Barnes-Hut t-SNE [68][2] to visualize the deeply learnt features of the samples from Market1501. It can be seen that softmax loss can learn a good deep embedding with large inter-class separation for person re-id. After combing softmax loss and center loss, the learnt features of the same person get a little closer to each other and D_c is reduced (16.72 versus 9.39). Besides, combining softmax loss and inter-center loss can make the samples of overlapped identities stay a little apart than using softmax loss individually, and D_{ic} is enlarged (29.34 versus 53.28). The experimental results and visualizations reveal the effectiveness of center loss and inter-center loss in boosting person re-id performance.

[2] We use the implementation in https://github.com/DmitryUlyanov/Multicore-TSNE.

Running time analysis. In Table 8, we compare the average feature extraction time per image and the average retrieval time per image of our method with four other existing methods, including IDE(R) [55], CNN-Embedding [13], IDLA [33], and RCN [69]. For fair comparison, we re-implement their feature extraction codes using PyTorch. The experiments are conducted on a machine with Intel Xeon CPU (64G memory) and four Titan GPUs (48G memory in total). During retrieval phase, a batch is composed of 2,000 images. Our method, IDE(R) [55], and CNN-Embedding [13] use feature vectors that can be saved in buffer for distance calculation, and they only need to forward all the images once. But for IDLA [33] and RCN [69], they have to forward the same image for several times to obtain the joint feature of image pair. Therefore, our method, IDE(R) [55], and CNN-Embedding [13] exhibit high computation efficiency compared to IDLA [33] and RCN [69].

5. Conclusion

In this paper, we present a hybrid deep model that combines multiple loss functions for person re-id task. The multi-loss function consists of softmax loss, center loss, and inter-center loss. Softmax loss enforces the separation between different person identities, while center loss reduces intra-class differences, and meanwhile, inter-center loss pushes apart different identity centers. Through the combination strategy, our method can learn more discriminative pedestrian descriptors with inter-class separation and intra-class compactness. We conduct extensive experiments on three benchmark person re-id datasets including Market-1501, CUHK03, and DukeMTMC-reID. The experimental results demonstrate that our method obtains better performance than most of state-of-the-art approaches.

Conflict of interest

There are no conflicts of interest.

Acknowledgments

This study is partially supported by National Natural Science Foundation of China under Grant 61673274, and Shanghai Science and Technology Commission Scientific Research Project with project Nos. 17DZ1100803.

References

[1] H. Shi, Y. Yang, X. Zhu, S. Liao, Z. Lei, W. Zheng, S.Z. Li, Embedding deep metric for person re-identification: A study against large variations, in: European Conference on Computer Vision, 2016, pp. 732–748.
[2] H. Liu, J. Feng, M. Qi, J. Jiang, S. Yan, End-to-end comparative attention networks for person re-identification, IEEE Trans. Image Process. 26 (7) (2017) 3492–3506.
[3] D. Yi, Z. Lei, S. Liao, S.Z. Li, Deep metric learning for person re-identification, in: International Conference on Pattern Recognition, 2014, pp. 34–39.
[4] R.R. Varior, B. Shuai, J. Lu, D. Xu, G. Wang, A siamese long short-term memory architecture for human re-identification, in: European Conference on Computer Vision, 2016, pp. 135–153.
[5] F. Wang, W. Zuo, L. Lin, D. Zhang, L. Zhang, Joint learning of single-image and cross-image representations for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1288–1296.
[6] D. Cheng, Y. Gong, S. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1335–1344.
[7] K. Sohn, Improved deep metric learning with multi-class n-pair loss objective, in: Neural Information Processing Systems, 2016, pp. 1857–1865.
[8] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: IEEE International Conference on Computer Vision, 2016, pp. 1116–1124.
[9] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3774–3782.
[10] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: IEEE International Conference on Computer Vision, 2017, pp. 3820–3828.
[11] D. Cheng, Y. Gong, W. Shi, S. Zhang, Person re-identification by the asymmetric triplet and identification loss function, Multimedia Tools Appl. 77 (3) (2017) 3533–3550.
[12] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camera style adaptation for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5157–5166.
[13] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person re-identification, Acm Trans. Multimedia Comput. Commun. Appl. 14 (1) (2017).
[14] L. Zheng, Y. Yang, A.G. Hauptmann, Person re-identification: Past, present and future, arXiv preprint arXiv: 1610.02984.
[15] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, 2016, pp. 499–515.
[16] Y. Yang, J. Yang, J. Yan, S. Liao, D. Yi, S.Z. Li, Salient color names for person re-identification, in: European Conference on Computer Vision, 2014, pp. 536–551.
[17] F. Xiong, M. Gou, O. Camps, M. Sznaier, Person re-identification using kernel-based metric learning methods, in: European Conference on Computer Vision, 2014, pp. 1–16.
[18] S. Khamis, C.H. Kuo, V.K. Singh, V.D. Shet, L.S. Davis, Joint learning for attribute-consistent person re-identification, in: European Conference on Computer Vision, 2014, pp. 134–146.
[19] M. Koestinger, M. Hirzer, P. Wohlhart, P.M. Roth, H. Bischof, Large scale metric learning from equivalence constraints, in: IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 2288–2295.
[20] W. Li, X. Wang, Locally aligned feature transforms across views, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3594–3601.
[21] K.Q. Weinberger, L.K. Saul, Distance metric learning for large margin nearest neighbor classification, J. Machine Learn. Res. 10 (1) (2009) 207–244.
[22] S. Liao, Y. Hu, X. Zhu, S.Z. Li, Person re-identification by local maximal occurrence representation and metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 2197–2206.
[23] D. Gray, H. Tao, Viewpoint invariant pedestrian recognition with an ensemble of localized features, in: European Conference on Computer Vision, 2008, pp. 262–275.
[24] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, Person re-identification by symmetry-driven accumulation of local features, in: IEEE Conference on Computer Vision and Pattern Recognition, 2010, pp. 2360–2367.
[25] I. Kviatkovsky, A. Adam, E. Rivlin, Color invariants for person reidentification, IEEE Trans. Pattern Anal. Mach. Intell. 35 (7) (2013) 1622–1634.
[26] Y. Duan, J. Lu, J. Feng, J. Zhou, Context-aware local binary feature learning for face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (5) (2018) 1139–1153.
[27] J. Lu, V.E. Liong, J. Zhou, Simultaneous local binary feature learning and encoding for homogeneous and heterogeneous face recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (8) (2017) 1979–1993.
[28] J.V. Davis, B. Kulis, P. Jain, S. Sra, I.S. Dhillon, Information-theoretic metric learning, in: International Conference on Machine Learning, 2007, pp. 209–216.
[29] W.-S. Zheng, S. Gong, T. Xiang, Reidentification by relative distance comparison, IEEE Trans. Pattern Anal. Mach. Intell. 35 (3) (2013) 653–668.
[30] S. Pedagadi, J. Orwell, S. Velastin, B. Boghossian, Local fisher discriminant analysis for pedestrian re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3318–3325.
[31] M. Hirzer, P.M. Roth, H. Bischof, Person re-identification by efficient impostor-based metric learning, in: International Conference on Advanced Video and Signal-Based Surveillance, 2012, pp. 203–208.
[32] S.-Z. Chen, C.-C. Guo, J.-H. Lai, Deep ranking for person re-identification via joint representation learning, IEEE Trans. Image Process. 25 (5) (2016) 2353–2367.
[33] E. Ahmed, M. Jones, T.K. Marks, An improved deep learning architecture for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3908–3916.
[34] L. Wu, C. Shen, A. v. d. Hengel, Personnet: Person re-identification with deep convolutional neural networks, arXiv preprint arXiv: 1601.07255.
[35] R.R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision, 2016, pp. 791–808.
[36] J. Lu, J. Hu, J. Zhou, Deep metric learning for visual understanding: An overview of recent advances, IEEE Signal Process. Mag. 34 (6) (2017) 76–84.
[37] J. Lu, J. Hu, Y.-P. Tan, Discriminative deep metric learning for face and kinship verification, IEEE Trans. Image Process. 26 (9) (2017) 4269–4282.
[38] J. Hu, J. Lu, Y.-P. Tan, Sharable and individual multi-view metric learning, IEEE Trans. Pattern Anal. Mach. Intell. 40 (9) (2018) 2281–2288.
[39] Y. Duan, W. Zheng, X. Lin, J. Lu, J. Zhou, Deep adversarial metric learning, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2780–2789.
[40] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person re-identification, arXiv preprint arXiv: 1703.07737.
[41] T. Xiao, H. Li, W. Ouyang, X. Wang, Learning deep feature representations with domain guided dropout for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 1249–1258.
[42] Z. Zhong, L. Zheng, Z. Zheng, S. Li, Y. Yang, Camstyle: A novel data augmentation method for person re-identification, IEEE Trans. Image Process. 28 (3) (2019) 1176–1190.
[43] J.Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2242–2251.
[44] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Y. Yang, Improving person re-identification by attribute and identity learning, arXiv preprint arXiv: 1703.07220.
[45] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1077–1085.
[46] S.E. Wei, V. Ramakrishna, T. Kanade, Y. Sheikh, Convolutional pose machines, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4724–4732.
[47] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person re-identification, arXiv preprint arXiv: 1701.07732.
[48] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: International Joint Conference on Artificial Intelligence, 2017, pp. 2194–2200.
[49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[50] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Neural Information Processing Systems, 2012, pp. 1097–1105.
[51] W. Li, R. Zhao, T. Xiao, X. Wang, Deepreid: Deep filter pairing neural network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 152–159.
[52] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.
[53] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, in: European Conference on Computer Vision, 2016, pp. 17–35.
[54] H. Moon, P.J. Phillips, Computational and performance aspects of pca-based face-recognition algorithms, Perception 30 (3) (2001) 303–321.
[55] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1318–1327.

[56] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[57] H. Zhao, M. Tian, S. Sun, J. Shao, J. Yan, S. Yi, X. Wang, X. Tang, Spindle net: Person re-identification with human body region guided feature decomposition and fusion, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 907–915.
[58] M.S. Sarfraz, A. Schumann, A. Eberle, R. Stiefelhagen, A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking, in: IEEE International Conference on Computer Vision, 2018, pp. 420–429.
[59] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2018) 1.
[60] W. Chen, X. Chen, J. Zhang, K. Huang, Beyond triplet loss: a deep quadruplet network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 403–412.
[61] W. Li, X. Zhu, S. Gong, Harmonious attention network for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 2285–2294.
[62] J. Si, H. Zhang, C.G. Li, J. Kuen, X. Kong, A.C. Kot, G. Wang, Dual attention matching network for context-aware feature sequence based person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5363–5372.
[63] T. Xiao, S. Li, B. Wang, L. Lin, X. Wang, Joint detection and identification feature learning for person search, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3376–3385.
[64] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017, pp. 1435–1443.
[65] R. Yu, Z. Dou, S. Bai, Z. Zhang, Y. Xu, X. Bai, Hard-aware point-to-set deep metric for person re-identification, in: European Conference on Computer Vision, 2018, pp. 188–204.
[66] M. Jaderberg, K. Simonyan, A. Zisserman, et al., Spatial transformer networks, in: Neural Information Processing Systems, 2015, pp. 2017–2025.
[67] G. Huang, Z. Liu, L.V.D. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2261–2269.
[68] L.V.D. Maaten, Accelerating t-sne using tree-based algorithms, J. Machine Learn. Res. 15 (1) (2014) 3221–3245.
[69] W. Zhong, L. Jiang, T. Zhang, J. Ji, H. Xiong, Combining multilevel feature extraction and multi-loss learning for person re-identification, Neurocomputing 334 (21) (2019) 68–78.