Knowledge-Based Systems 187 (2020) 104832

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

Effective person re-identification by self-attention model guided feature learning☆

Yang Li [a,1], Xiaoyan Jiang [a,*,1], Jenq-Neng Hwang [b]

[a] School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, No., 333 of Longteng Road, Shanghai, China
[b] Department of Electrical and Computer Engineering, University of Washington, Box 352500, Seattle, WA 98195, USA

ARTICLE INFO

Article history:
Received 22 January 2019
Received in revised form 30 June 2019
Accepted 3 July 2019
Available online 8 July 2019

Keywords:
Person re-identification
Feature extraction
Self-attention
Cross-entropy loss
Triplet loss

ABSTRACT

Person re-identification (re-ID), of which the goal is to recognize person identities of images captured by non-overlapping cameras, is a challenging topic in computer vision. Most existing person re-ID methods conduct directly on detected objects, which ignore the space misalignment caused by detectors, human pose variation, and occlusion problems. To tackle the above mentioned difficulties, we propose a self-attention model guided deep convolutional neural network(DCNN) to learn robust features from image shots. Kernels of the self-attention model evaluate weights for the importance of different person regions. To solve the local feature dependence problem of feature extraction, the non-local feature map generated by the self-attention model is fused with the original feature map generated from the resnet-50. Furthermore, the loss function considers both the cross-entropy loss and the triplet loss in the training process, which enables the network to capture common characteristics within the same individuals and significant differences between distinct persons. Extensive experiments and comparative evaluations show that our proposed strategy outperforms most of the state-of-the-art methods on standard datasets: Market-1501, DukeMTMC-reID, and CUHK03.

1. Introduction

Person re-ID has various important application scenarios, for example, video surveillance, human behavior analysis, and multi-target tracking. The goal of person re-ID is to retrieve specific persons in images or video sequences obtained by multiple non-overlapping cameras. Given a query person's image, the task is to find correct corresponding matches among a set of candidate images captured by different cameras in the gallery. Two crucial issues are needed to be addressed: reliable image feature extraction is required to represent both the query and the gallery images, and suitable distance metric is indispensable to determine whether a gallery image represents the same individual as the query image. Person re-ID is challenging because of the following difficulties:

First of all, due to the changing of camera viewpoint and the variability of the pedestrian's posture, the detected body parts of the person in two images are misaligned. Sample detected bounding boxes of the same person are shown in Fig. 1(a). The body parts labeled as green areas of two images should be aligned, but spatially the left green region of the human head corresponds to the right red area of the background. As a result, feature maps of the two regions extracted by convolutional neural networks (CNNs) are significantly different and cannot be directly compared. Even when the detection is accurate, there still exists the spatial misalignment problem due to the varying of human poses. As shown in Fig. 1(b), the right leg region of the left image (green box) is spatially misaligned with the left leg region (red box) in the right image. Moreover, the black bag cannot be observed in the left image which causes changes in appearance characteristics. To extract reliable features for person re-ID, there are different feature extraction methods. For example, features can be directly extracted [1] from the whole image, the image can be divided into several regions that are then extracted for matching [2], or the spatial alignment method is directly proposed [3]. However, the features cannot be well aligned and are not robust to detection errors or human pose variations. In this paper, we adopt the CNN resnet-50 [4], which is good at preventing gradients from disappearing in deep-layer networks, to extract the features of the whole image, and introduce the model of attention mechanism to solve the problem of local feature dependency and space misalignment between image pairs.

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2019.07.003.

* Corresponding author.
E-mail address: xiaoyan.jiang@sues.edu.cn (X. Jiang).

[1] Yang Li and Xiaoyan Jiang contribute equally to this work.

https://doi.org/10.1016/j.knosys.2019.07.003
0950-7051/

Fig. 1. Challenges of person re-ID: (a) misalignment due to inaccurate detection, (b) misalignment due to pose changes, (c) appearance in consistency, (d) occlusion.

Moreover, different people may wear alike or have similar gestures. As shown in Fig. 1(c), the general appearances of two different persons are quite similar. In this case, feature extraction focusing on salient differences and fine-grained details between different categories is helpful for further feature matching. But most existing person re-ID methods have difficulties to capture such detailed information. In this paper, we present a multi-scale spatial attention model and combine the triplet loss with the cross-entropy loss to obtain robust features for each person, resulting in a network that can better distinguish inter-person and intra-person.

Lastly, occlusions between objects appear in most practical scenarios. Some body parts may be occluded by scene objects or other people in the image, which make the identifying more difficult. As shown in Fig. 1(d), the middle region of the pedestrian in the right image is occluded by a blue sign. Thus the feature importance of this region should be decreased in the feature matching process. Due to the lack of partial body area information, we pay more attention to the correlation information of the rest parts of the same person in our network. Each pedestrian is treated as a category. We adopt the softmax cross-entropy loss function to train the classification tasks, ensuring that the same individuals' discriminant features can be obtained for better matching.

In order to obtain robust and discriminant features specifically for person re-ID, we propose a self-attention model based feature extraction framework. Distinct convolution kernels are combined with non-linear functions to get multi-scale spatial features, which can represent images more accurately. Furthermore, different loss items are fused in the loss function to preserve inter-person similarity and enlarge intra-person distinction. Our approach achieves the state-of-the-art performance on most standard datasets evaluated by the standard metric. The main contributions are summarized as the following two-folds:

- A multi-scale attention model is jointly trained with resnet-50 to solve the spatial misalignment and local feature dependency problems of person re-ID. The nonlinear combination of features from multiple scales in the self-attention mechanism merges the global and local features of the image effectively.
- The cross-entropy loss function usually used in multi-classification task is combined with the triplet loss function for person re-ID in the supervised training process of the network. The trained model extracts the similar characteristics of the same individuals better and significantly enlarges feature distinctions between different persons.

The remaining of this paper is organized as follows. We review the related work on person re-ID in Section 2. In Section 3, we present our approach, the network architecture, and the specific details of each algorithmic module. Experimental results and analysis are shown in Section 4. Section 5 summarizes our work and gives future plans.

2. Related work

To overcome the challenges discussed above, there have been numerous research efforts on person re-ID, which can be broadly categorized into the following techniques: representation learning, metric learning, local feature-based, generative adversarial networks, unsupervised learning-based, video sequence-based.

Representation learning. Representation learning is a commonly used person re-ID method [5–7], most of which belong to supervised learning relying on feature extraction methods . Since CNN automatically extracts apparent features from original image data according to the task requirements, some researchers regard person re-ID as a classification or a verification problem: (a) using the pedestrian's ID or attribute as label to train the model; (b) verification is meant to input a pair of pedestrian images so that the network has the ability to identify whether the two images belong to the same pedestrian or not. In the paper [5], the network includes a classification network and a verification network. The former performs ID prediction, and the latter combines the characteristics of two images to determine whether they belong to the same pedestrian. After sufficient data training, the network automatically extracts the feature for person re-ID when input a test image. However, authors in [6] state that the holistic pedestrian ID is not enough to learn a model with strong generalization ability, some of them additionally label the attribute characteristics of the pedestrian image, such as hair, gender, clothing, etc. As described in the paper [6], they improve the generalization ability of the network via combining ID loss and other attribute loss.

Metric learning. Metric learning is widely used in the field of image retrieval. Unlike representation learning, metric learning aims to learn the similarity between two images through the learned network. In other words, for the specific task of person re-ID, the similarity of different images from the same person is definitely greater than the different images from different pedestrians. That is to say, the loss function of the network ensures the distance of the same pedestrian images (positive sample pairs) are as small as possible, and the distance of different pedestrian images (negative sample pairs) are as large as possible. Common used loss methods for metric learning are contrastive loss [8], triplet loss [9], quadruplet loss [10], triplet hard loss with batch hard mining [11]. Among them, contrastive loss is mostly used to train a siamese network, which differentiates two inputs. Triplet loss function is a widely used in metric learning methods. The metric learning can be optimized on the basis of the performance of various loss functions.

Local feature-based. Global features got much attention in most early person re-ID works. People gradually discovered that the global feature encountered a bottleneck, they begin to learn the local features. Commonly used ideas for extracting local features include image dicing, skeleton key point localization, and pose correction. Among them, image dicing is a common way to extract local features. In the paper [12], an image is divided into several blocks vertically, and the features of each image block are

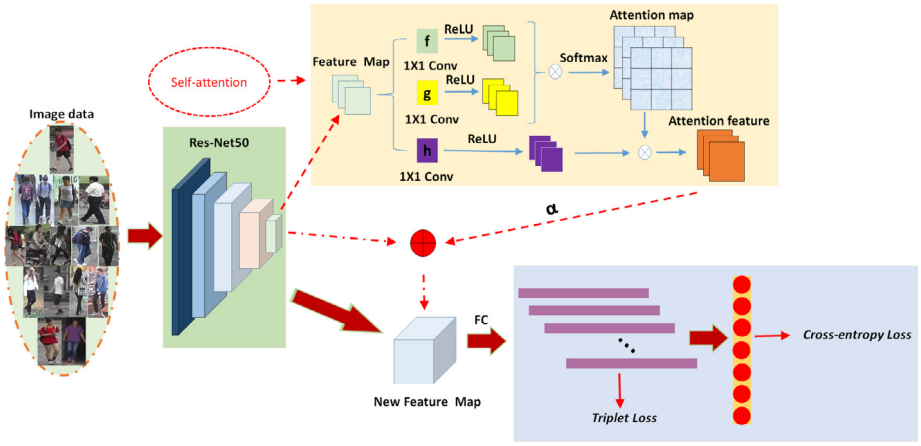


Fig. 2. The proposed network architecture. It consists of three parts: the restnet-50 network, the self-attention model, and the loss function. The self-attention model includes multi-scale transformation spaces and nonlinear combination of feature maps. The loss function considers both intra-person similarity and inter-person distinction.

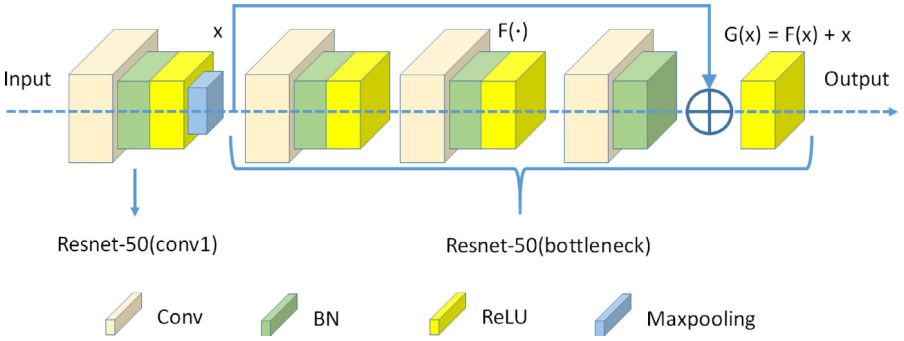


Fig. 3. Blocks of the resnet-50 architecture: the resnet-50 (conv1) basic frame of the first convolution layer and the resnet-50 (bottleneck) basic frame of res2c to res5c. In the resnet-50 (bottleneck), X is input, F is the conversion function, and G is output.

extracted and merged respectively. However, the disadvantage of this method is that the alignment of the image is relatively simple and straightforward, and it is easy to make the model judged incorrectly. In order to solve this problem, prior knowledge is used to first spatially align pedestrians, such as pre-trained human poses and models of skeleton key points. In the paper [13], the pose estimation model is used to estimate the key points of pedestrians. Then, the affine transformation is used to align the same key points. The key points are used to divide the human body into several regions and different scale local features are extracted for alignment. Similarly, spindle net [14] also makes use of the key points of the human body, but the difference is that they use these key points to get the body region of interest (ROI), from which the global features of the whole image and the multi-scale local features of each ROI can be fused for person re-ID. The paper [15] proposed a global–local-alignment descriptor to solve the problem of pedestrian pose transformation. Comparing to spindle net, the difference is that each part can calculate its individual loss, not a total loss calculated by merging the feature.

Generative adversarial networks. The acquisition of person re-ID dataset is another challenge. At present, the largest person re-ID dataset contains thousands of IDs, resulting a total number of tens of thousands of images. After the emergence of the generative adversarial networks (GAN), the model can be used to expand the dataset as much as possible to improve the learning ability of the network. The paper [16] is the first article to use GAN for person re-ID data augmentation. Even though the quality of the generated images is not high, but because of the enhanced dataset and the method of label smoothing regularization for outliers, the person re-ID performance is significantly enhanced. Subsequently, a new method on from the same group [17] is proposed to translate the image of one camera to another camera so that the network learns the differences of two cameras under different viewpoints. Based on this idea, the paper [18] uses GAN to migrate pedestrians from one dataset to another, which can effectively reduce the impact of different datasets due to environmental differences and the loss of network convergence. Moreover, difference pedestrian pose causes more difficulties to the task in hand. Thus, authors in [19] uses GAN to generate a series of standard pose images, which extracted a total number of eight poses covering all of the viewing angles. The final feature then combines information from all different poses to mitigate the problems caused by the scenario of varying poses.

Unsupervised learning-based. Supervised learning-based person re-ID methods require and rely on a large number of labeled information. Unsupervised learning methods [20–22] are proposed to learn cross-view identified information using unlabeled data. The paper [23] has achieved certain effects by transferring the visual invariant features of the characteristics from the source tag dataset to the unlabeled target dataset through the dictionary learning mechanism. The paper [24] proposes a new deep association learning (DAL) method to learn the model by jointly optimizing two interval-based association losses in an end-to-end manner, which effectively limits the association between each frame and the best matching of the same camera representation. The paper [25] proposed a new method of spatial–temporal information fusion based on Bayesian algorithm to achieve obvious improvement. Nonetheless, because an unsupervised learning lacks prior knowledge, its effect is in general not as good as supervised learning.

Video sequence-based. With the availability of videos collected from multiple connected cameras, people start to perform the task of person re-ID on the video sequences, which not only contains the spatial content information of each individual image, but also the temporal motion information in the video that can be jointly taken into account. The most common idea in dealing with video sequences is to use recurrent neural networks (RNNs), where each frame of the image can be combined with the spatial feature of the frame and the motion characteristics between frames. In the paper [26], when there is occlusion in a single-frame image, information from other frames can be used to compensate. Due to the development of video decoder and encoder, the attention model has gradually been valued and widely quoted, such as [27], where a new joint learning multi-granular attention selection and feature representation method is proposed to optimize the person re-ID problem. In the paper [28], the fusion of temporal and spatial attention mechanism effectively improves person re-ID results.

3. The proposed approach

As shown in Fig. 2, we propose a deep learning framework to deal with image-based person re-ID, where image shots are sequentially fed to a pre-trained resnet-50 network (Section 3.1) for feature extraction to obtain corresponding feature maps. The feature map is then input to the self-attention network (Section 3.2) to generate new discriminative spatial visual features. Finally, the new feature map is flattened into feature vectors, which are sent to the fully connected layer for supervised learning of the entire network. To better represent intra-person features and distinguish inter-person differences, the loss function fuses the cross-entropy with the triplet loss (Section 3.3) to give informative feedback to the network. Details of the proposed framework are described as the following.

3.1. Baseline network resnet-50

Nowadays, there are many convolutional neural network (CNN) architectures, among which resnet-50 is shown to prevent the gradient from disappearing and obtain more representative features. Thus, we used the resnet-50 CNN [4] as our basic network framework to extract spatial visual features. Sample images are cropped to the size of 256 × 128, including rotation and translation of the image data augmentation. Then they are used as the training set for our CNN. According to the conventional fine-tuning strategy, we use the pre-trained model on the ImageNet dataset [29], of which we only reserve the convolution layer by removing the fully connected layer. The simplest residual network block layer is used to reduce the amount of calculation and highlight the effect of the proposed attention model.

The resnet-50 network contains an independent convolution layer (conv1) at the forefront, which is followed by four residual blocks (from res2c to res5c). The basic structure of resnet-50 and the corresponding parameter configuration of each module are shown in Fig. 3 and Table 1, respectively. BN refers to batch normalization and the ReLU is used as the activation function. Conv1 and res2c to res5c are used for feature extraction. Consequently, each image I_n is represented by a feature map $\{f_{n,l}\}_{l=1,\cdots,L}$ composed by 8 × 4 grids, where $L = 32$ is the number of batch size. Each feature map is a $D = 2048$ dimensional vector.

Table 1
Baseline network of the resnet-50 model.

Layer		Output size
Conv 1	7 × 7, 64, stride 2,2	64 × 128 × 64
Pool 1	3 × 3, max, stride 2,2	64 × 64 × 32
Res2c	$\begin{bmatrix} 1\times1, 64 \\ 3\times3, 64 \\ 1\times1, 256 \end{bmatrix} \times 3$	256 × 64 × 32
Res3c	$\begin{bmatrix} 1\times1, 128 \\ 3\times3, 128 \\ 1\times1, 512 \end{bmatrix} \times 4$	512 × 32 × 16
Res4c	$\begin{bmatrix} 1\times1, 256 \\ 3\times3, 256 \\ 1\times1, 1024 \end{bmatrix} \times 6$	1024 × 16 × 8
Res5c	$\begin{bmatrix} 1\times1, 512 \\ 3\times3, 512 \\ 1\times1, 2048 \end{bmatrix} \times 3$	2048 × 8 × 4

3.2. Self-attention model

The generalized attention mechanism is mostly used in the process of sequence encoder and decoder. As shown in [31], scaled dot-product attention function conducts on a set of queries simultaneously and packs the formation together into a matrix Q. The Keys and Values are also packed together into the matrices K and V, respectively. The attention function is defined as:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt[2]{d_k}})V, \tag{1}$$

where $Q \in \mathbb{R}^{n\times d_k}$, $K \in \mathbb{R}^{m\times d_k}$, $V \in \mathbb{R}^{m\times d_v}$, and d_k is the scale factor. The softmax function is used for calculating weight values.

Refer to the above attention mechanism, a novel self-attention model is proposed as shown in Fig. 2. To obtain scale-invariant attention features, different transformation spaces are created in the model to obtain features of different scales on the original feature space. Moreover, ReLU is combined to get a nonlinear response. Each point in the feature map is activated with a corresponding weight value, which solves the spatial misalignment and local feature dependency problems.

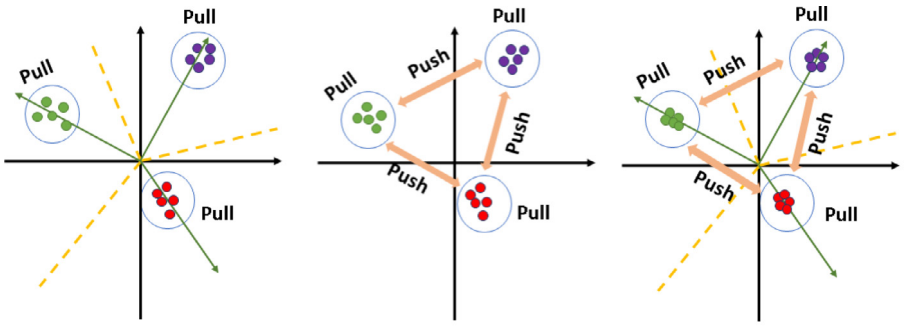
We compute the visual spatial feature of the image $\{f_{n,l}\}_{l=1,\cdots,L}$ by resnet-50, which is simplified to $x \in \mathbb{R}^{D\times L}$ while keeping the size as the original feature map size (h, w). Then the feature x is converted by two different feature transformation space function f and g to calculate the attention score, and another different transformation space function h to generate a new feature map. Thus, we have three different feature transformation spaces $k \in (f, g, h)$. We indicate $W_f \in \mathbb{R}^{\bar{D}\times D}$, $W_g \in \mathbb{R}^{\bar{D}\times D}$, $W_h \in \mathbb{R}^{\bar{D}\times D}$ to be parameter matrices trained by the network. Dot-product with attention weight is defined by the following formula:

$$f(x) = W_f x, g(x) = W_g x, h(x) = W_h x. \tag{2}$$

Here, we use 1 × 1 convolution to perform linear combination of feature maps in different scales and different spaces. Since the size of the original feature map is large, the dot-product requires a large memory size. We convert them to a lower-dimensional feature space $\bar{D} = D/8$ by filtering the original feature map using a small number of kernels.

After linear space conversion, the ReLU activation function is applied to perform nonlinear processing to obtain the response $e_{i,j}$, which is corresponding to the original feature map. The characteristic response transformation formula is:

$$e_{i,j} = (max(f(x_i), 0))^T (max(g(x_i), 0)). \tag{3}$$



(a) cross-entropy loss (b) triplet loss (c) cross-entropy + triplet loss

Fig. 4. Visualization of the sample distribution in the embedding space supervised by (a) cross-entropy loss, (b) triplet loss, and (c) cross-entropy + triplet loss. Colored points represent embedding features from distinct persons. The assumed person identification hyperplanes are drawn by yellow dotted lines [30].

Table 2
distribution of Market-1501, DukeMTMC-reID and CUHK03.

Dataset	Cameras	Train		Test			
				query		gallery	
		People	Images	People	Images	People	Images
Market-1501	6	751	12936	750	3368	750	19732
DukeMTMC-reID	8	702	16522	702	2228	1110	17661
CUHK03	2	767	7365	700	1400	700	5332

Based on the obtained response $e_{i,j}$, our attention score $s_{j,i}$ is calculated as follows:

$$s_{j,i} = \frac{exp(e_{i,j})}{\sum_{i=1}^{D} exp(e_{i,j})}, \tag{4}$$

where $s_{j,i}$ denotes the extent to which the model attends to the ith location on the feature map when synthesizing the jth dimension. By definition, each receptive field is a probability mass function since $\sum_{i=1}^{L}\sum_{j=1}^{D} s_{i,j} = 1$.

We denote the output of the attention mechanism layer as $O = (o_1, o_2, \cdots, o_j, \cdots, o_N) \in \mathbb{R}^{D\times L}$, where o_j is defined as

$$o_j = \sum_{i=1}^{L}\sum_{j=1}^{D} s_{i,j} h(x). \tag{5}$$

In addition, we further multiply the output of the attention layer by a scale parameter α, of which the result is added back to the input feature map. Therefore, the final output y is given by

$$y = \alpha o_j + x_i, \tag{6}$$

where α is initialized as zero and is gradually updated. Such parameter setting makes the network to give priority to local neighborhoods at the beginning and then gradually learn to assign high weights to non-local evidence.

In contrast to spatial attention [27], the self-attention model directly conducts on the feature map. For each feature point, the nonlinear function is activated to obtain the weight value of each spatial points in the feature map. The weight value determines the effect of the corresponding feature point on the entire task. It combines the diversity of global features and local features to solve the problem of over-reliance on local features.

3.3. Loss function

We combine the triplet loss function with hard mining [11] and the softmax cross-entropy loss function with label smoothing regularization [32] in the training process of the CNN network. The triplet loss forces the network to learn robust features representing key traits of the same person in different views, while keeping the differences between different pedestrians. In addition, the softmax cross-entropy loss facilitates the learning of representative features expressing common characteristics of the same individuals in different scenes. As shown in Fig. 4(a), cross-entropy loss tries to separate the embedding space into different subspaces by hyperplanes. In contrast, as shown in Fig. 4(b), triplet loss aims to pull intra-class similarity and push inter-class discrimination. Thus, these two losses are combined to learn more effective and discriminative features as shown in Fig. 4(c).

The triplet loss function is originally proposed in [11], which is named as Batch Hard triplet loss function. For each batch, we randomly select the number of P different individuals in the dataset. For each person, K tracks are randomly selected. A single track is composed by images captured from different viewing angles or extracted from different sequences. Since the proposed person re-ID approach needs only images not videos, each track here can contains only one image. We set $T = 1$ and each batch contains $P \times K$ images in total.

For each sample a in the batch, the hardest positive and the hardest negative samples within the batch are selected to form the triplet for computing the loss $L_{triplet}$:

$$L_{triplet} = \overbrace{\sum_{i=1}^{p}\sum_{a=1}^{k}}^{\text{all anchors}} [m + \overbrace{\max_{p=1\cdots K} D(f_a^i, f_p^i)}^{\text{hardest positive}} - \overbrace{\min_{\substack{j=1\cdots P \\ n=1\cdots K \\ j\neq i}} D(f_a^i, f_n^i)}^{\text{hardest negative}}]_+. \tag{7}$$

The original softmax cross-entropy loss function is given by:

$$L_{softmax} = -\frac{1}{P\times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} log q_{i,a}, \tag{8}$$

where $p_{i,a}$ is the ground truth identity and $q_{i,a}$ is the prediction of the sample $\{i, a\}$. The label-smoothing regularization is used to regularize the model resulting in

$$L'_{softmax} = -\frac{1}{P\times K}\sum_{i=1}^{P}\sum_{a=1}^{K} p_{i,a} log((1-\varepsilon)q_{i,a} + \frac{\varepsilon}{N}), \tag{9}$$

which is a mixture of the original ground-truth distribution $q_{i,a}$ and the uniform distribution $u(x) = \frac{1}{N}$ with weights 1-ε and ε, respectively. We set ε = 0.1 and N is the number of classes.

The total loss L is the combination of the above mentioned two loss items:

$$L = L_{triplet} + L'_{softmax}. \tag{10}$$

The two loss items complement and constrain each other to ensure that the model not only has a good generalization ability but also prevents the model from over-fitting. The loss value

Table 3
Ablation study using the single query.

Database	resnet-50	Spatial-attention	Ours	Evaluation				
				Rank-1	Rank-5	Rank-10	Rank-20	mAP
Market-1501	√	×	×	86.8%	94.0%	96.8%	98.0%	75.6%
	√	√	×	88.6%	93.9%	97.1%	98.0%	76.2%
	√	×	√	**90.2%**	96.7%	98.1%	99.0%	**82.7%**
DukeMTMC-reID	√	×	×	76.8%	88.0%	90.2%	91.6%	74.2%
	√	√	×	78.2%	89.6%	92.4%	92.3%	75.8%
	√	×	√	**81.0%**	92.4%	94.2%	95.9%	**78.0%**
CUHK-03(labeled)	√	×	×	46.9%	72.0%	80.5%	86.4%	58.2%
	√	√	×	58.6%	76.3%	85.4%	88.7%	62.8%
	√	×	√	**64.8%**	84.3%	89.3%	93.2%	**72.7%**
CUHK-03(detected)	√	×	×	45.6%	70.2%	78.6%	85.4%	55.3%
	√	√	×	56.5%	74.8%	83.2%	88.5%	60.2%
	√	×	√	**60.5%**	80.1%	87.4%	91.8%	**67.8%**

is iteratively updated to prevent the gradient from disappearing, which makes the network more stable in the convergence process.

4. Experiments

Our approach is evaluated on three standard datasets: Market-1501[33] and Duke MTMC-reID [34], which are collected by multiple cameras and have relatively large-scale sample images. In addition, we show the experimental results on the CUHK03 [35] dataset, which has a small number of images. The dataset is separated into the training set, the query set, and the gallery set to achieve end-to-end training and testing. The experimental evaluation demonstrates that our proposed framework obtains better performance than most state-of-the-art methods.

4.1. Dataset

Market-1501. The Market-1501 dataset was recorded by six different cameras. It consists of 32,668 images from 1501 people, which is modeled and deformed using the Deformable Part Model (DPM) [36]. The dataset is divided into three parts: 751 people with totally 12,936 images as the training set, 750 people with totally 19,732 images as the gallery, and 3,368 images selected from the same 750 people as the query. All images are with the size of 128 × 64.

DukeMTMC-reID. The DukeMTMC-reID dataset is a subset of the newly released multi-target multi-camera pedestrian tracking dataset. The original dataset contains eight 85-minutes' high-resolution videos captured from eight different cameras. Bounding boxes of pedestrians are manually labeled. Similar to the format of the Market-1501 dataset, 36,411 images of 1812 different pedestrians recorded from eight different viewing angles are obtained. The dataset is divided into three parts: 702 people with totally 16,522 images as the training set, 17,661 images of 1110 different people as the gallery. In addition, a total number of 2,228 images of 702 people from the initial selection of the gallery as the query. The size of the image is 128 × 64, which is the same as that of the market-1501 dataset.

CUHK03. The CUHK03 dataset contains 14096 images of 1467 different people, each of them is captured by two cameras in the campus of the Chinese University of Hong Kong. Each person has an average of 4.8 images per camera. It contains two image sets, one of which is manually labeled and the other is derived from the Deformable Part Model (DPM). We use 7365 images of different 767 people as the training set, and the remaining 700 people with total 5332 images as the gallery. In addition, the same people in the gallery have a total of 1400 images as the query set.

We summarize the data distributions and show several samples of the above mentioned three datasets in Table 2 and Fig. 5, respectively.

4.2. Evaluation criteria and parameter configuration

Experimental results are evaluated by the accuracy of rank-1, 5, 10, 20 and the mean average precision (mAP) metric. Rank-m (m=1,5,10,20) here indicates different hit probability. The mAP denotes the mean of different hit probabilities.

The resnet-50 network pre-trained on ImageNet is used to initialize the convolutional layer of the proposed network to extract image features. We add new convolution modules and full connected layers to construct our classification framework. The size of all images is set to be 256 × 128. In the process of data augmentation, all the images are rotated, then moved and flipped in both the horizontal and vertical directions. Finally, all the pixel values are normalized. Average values used for RGB channels are 0.485, 0.456, and 0.406, respectively, and the standard deviations are 0.229, 0.224, and 0.225, separately.

The model is trained for 800 epochs. The starting learning rate is 0.0003 and is reduced by the optimization algorithm Adam with a decay rate of 0.0005 for every 100 epochs. The batch size is 32. Images of one single person captured by the same camera are identified as one group. Each batch is randomly selected for training to prevent the problem of over-fitting training and low-speed convergence.

In the process of testing, the feature map of the last level of the attention model is expanded into a 2048-dimensional vector. Each image is read sequentially order from the query set. Similarly, features of images in the galley are extracted by the same network. The similarity of the two features is calculated by the Euclidean distance and is normalized to [0, 1].

4.3. Ablation studies

The experimental results of three different methods on the three different datasets are shown in Table 3. The first method is resnet-50-based without using the attention mechanism. The second uses spatial-attention and the third uses our proposed self-attention model. The spatial-attention method directly performs on a one-dimensional time-series convolution of each feature map to obtain the corresponding expression value, and each function value is transferred into a vector, which is substituted into the softmax function to obtain the weight value of each feature map. Finally, we fuse multiple feature maps with the last feature map for training based on the weight value.

As the experimental results show, the attention mechanism plays a certain role in the larger datasets Market-1501 and DukeMTMC-reID. The Rank-1 accuracy is improved by 3%–5%. While on the smaller dataset CUHK03, the Rank-1 accuracy is increased more than 10% by using the proposed attention model.

Fig. 5. Sample images from the datasets Market-1501, DukeMTMC-reID, and CUHK03.

Comparison with the state-of-the-art. As shown in Tables 4 and 5, we compare our results with the state-of-the-art methods on the dataset Market-1501 and DukeMTMC-reID, respectively. In the comparison, the metric learning based methods include Gated-Sia [8], Basel.+LSRO [16], DML [37], JLML [38], Basel.+OIM [39], and Verif.-Identif.+LSRO [16]. Deep learning based methods include Inception-V3[40], PDC [41], and Deep Transfer [5].

Figs. 6 and 7 show the curves of Rank-1 accuracy and the histogram of mAP with respect to the number of iterations on Market-1501 and DukeMTMC-reID datasets, respectively. The resnet-50 network has better performance in extracting 2D-image features, which can effectively prevent the gradient from disappearing, resulting in faster training convergence speed. We use the attention model to obtain different weight values based on the relationship between different feature maps. To obtain more representative feature map, it is better to combine different multi-scale feature map information is combined. Finally, our self-attention model expands the value of each point on the feature map and multiplies the matrix to make the response of the original response larger, and then calculates the corresponding weight value of each point to get more discriminative features, which effectively combine the local features and global features of the whole image.

4.4. Analysis and visualization

We perform the relevant comparison on the dataset Market-1501, and visualize the results for analysis. In the following section, parameter configuration of the work influences the experimental results.

Attention scores obtained from different activation functions influence the experimental results in the training process. Table 6 shows the experimental results by choosing either the softmax or the sigmoid function as the activation function. The former is often used in multi-classification tasks, of which the predicted results are normalized to [0, 1] and summed up to one. The latter is mostly used in binary classification tasks, of which the output is a value between [0, 1]. As can be seen from Table 6, the evaluation has an approximate 1% difference by using different activation functions and the network adopting the softmax function obtains higher performance. The activation function nonlinearize linear part of the network, effectively normalizing the values by probability values. The softmax function is consistently used for all our experiments.

In the training process, our loss function is composed by the cross entropy and the triplet loss. In order to access the effect of each loss, we did experiments by using only one of the two loss functions for person re-ID as shown in Table 7. As shown in Fig. 8, the curve of the loss value during training is smoothed using the logarithmic function. The cross entropy loss function mainly focuses on the common features within the class, without

Table 4
Comparison of state-of-the-art methods on the Market-1501 dataset.

Database	Market-1501	
Method	Rank-1	mAP
SCSP [42]	51.90	26.40
DNS [43]	61.02	35.68
Gated-Sia [8]	65.88	39.55
Spindle [14]	76.90	–
Basel.+LSRO [16]	78.06	56.23
PIE [13]	79.33	55.95
Verif.-Identif.[44]	79.51	59.87
DLPAR [33]	81.00	63.40
Deep Transfer [5]	83.70	65.50
Verif.-Identif.+LSRO [16]	83.97	66.07
PDC [41]	84.14	63.41
DML [37]	87.70	68.80
SSM [45]	82.20	68.80
PN-GAN [19]	89.43	72.58
Ours(Self-attention)	**90.20%**	**82.70%**

Table 5
Comparison of state-of-the-art methods on the DukeMTMC-reID dataset.

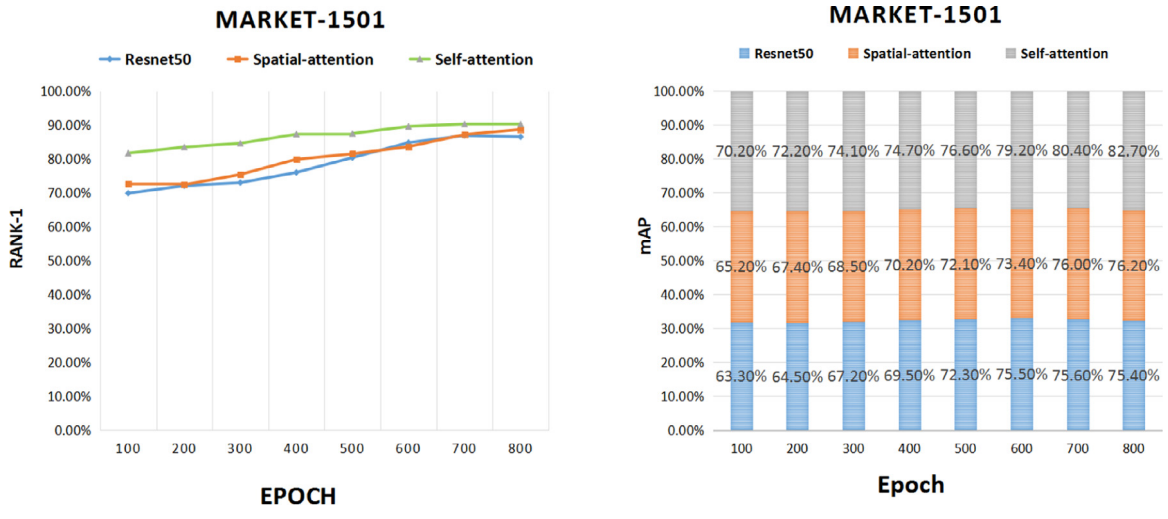
Database	DukeMTMC-reID	
Method	Rank-1	mAP
Basel.+LSRO [16]	67.70	47.10
Basel.+OIM [39]	68.10	–
AttIDNet [6]	70.69	51.88
ACRN [46]	72.60	52.00
SVDNet [47]	76.70	56.80
Chen et. al [48]	79.20	60.60
Inception-V3[40]	80.48	63.27
Ours(Self-attention)	**81.00%**	**78.00%**

Table 6
Evaluation on the dataset Market-1501 using different activation function.

Database	Activation function		Evaluation				
	softmax	sigmoid	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Market-1501	√		**90.2%**	96.7%	98.1%	99.0%	**82.7%**
		√	**89.4%**	95.9%	98.0%	99.0%	**81.5%**

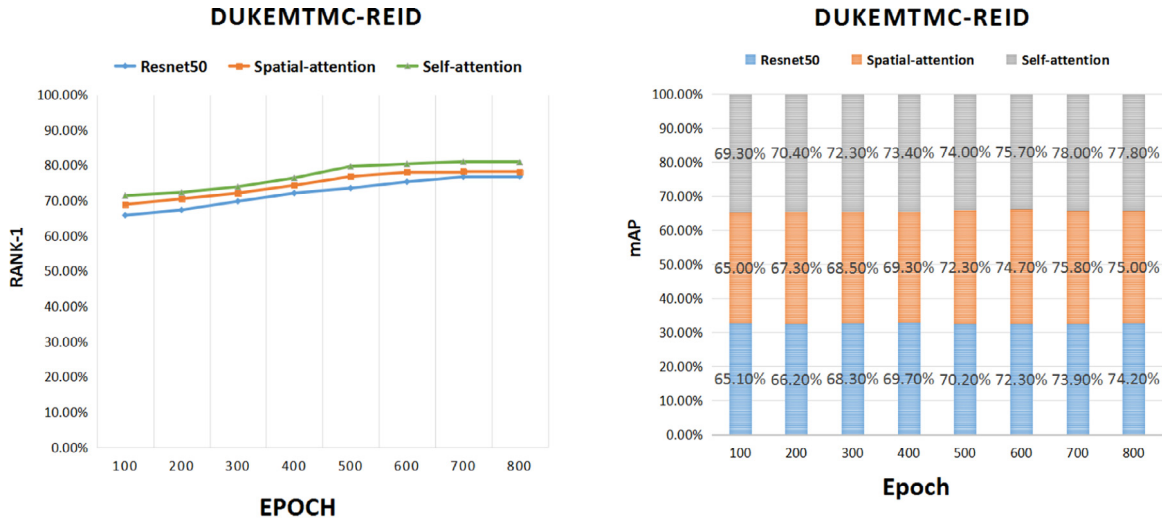
Table 7
Evaluation on the dataset Market-1501 using different loss function.

Database	Loss		Evaluation				
	Cross-Entropy	Triplet	Rank-1	Rank-5	Rank-10	Rank-20	mAP
Market-1501	√		85.3%	94.8%	97.0%	98.3%	76.4%
		√	86.2%	95.1%	97.3%	98.8%	76.8%
	√	√	**90.2%**	96.7%	98.1%	99.0%	**82.7%**



(a) (b)

Fig. 6. Evaluation on the dataset Market-1501: (a) Rank-1 curve, (b) mAP histogram.



(a) (b)

Fig. 7. Evaluation on the dataset DukeMTMC-reID: (a) Rank-1 curve, (b) mAP histogram.

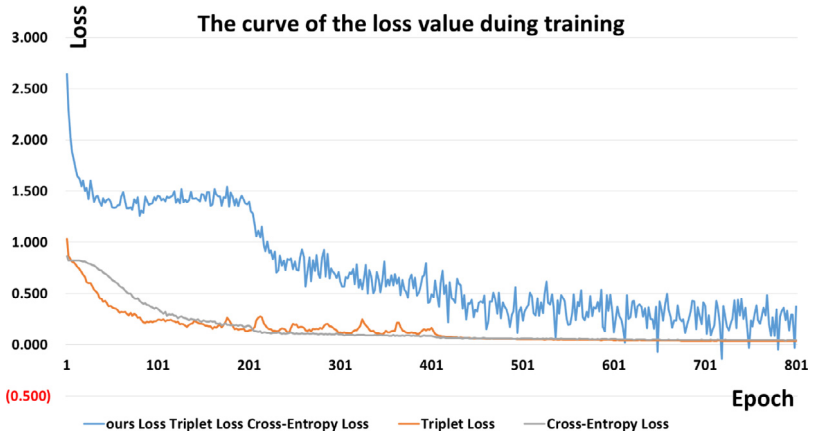


Fig. 8. Changes of loss values during training on the dataset Market-1501.

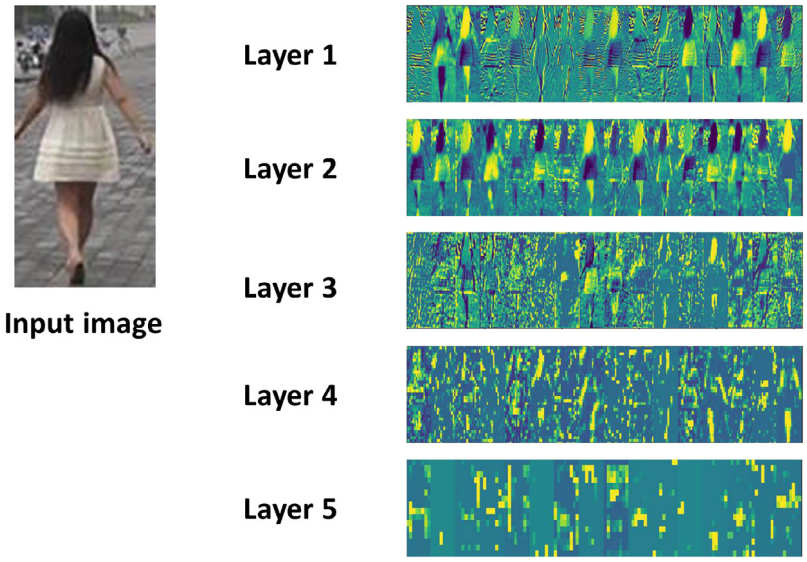


Fig. 9. Visualization of the feature maps of different layers in resnet-50.

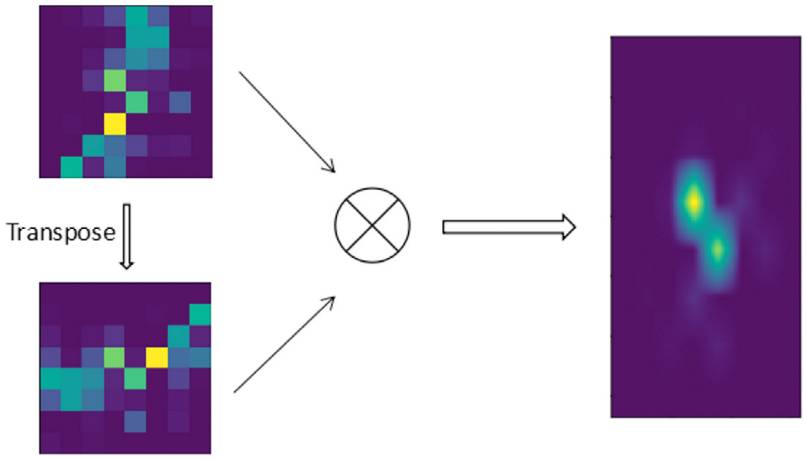


Fig. 10. Visualization of a kernel matrix (the left image) generated by the attention model. The corresponding attention model (the right image) is reshaped to the size of the resnet-50 feature map, which assigns higher weighting values to regions that are brighter.

paying attention to the differences with other classes, which is beneficial for person re-ID task to learn the characteristics of the same person in different conditions. In contrast, the triplet loss function considers both intra-class features and inter-class properties, helping the network obtaining more invariant features in the training process. Two loss functions are fused in our approach so that they are mutually constrained and interactive with each other. Moreover, the network convergences in a better performance, although in a slower influence of the speed.

To analyze the attention model in the network, features in the resnet-50 network layer as shown in Figs. 9, 10 shows an attention kernel, and Fig. 11 shows the resulting effect of the attention model on different pedestrians.

As can be seen in Fig. 9, features extracted by each layer of the network are unique. The low-level network extracts simple appearance features such as pedestrian contour, texture, color, and etc. The features extracted by high-level networks contain rich semantic information to make the final feature map more representative. Fig. 10 visualizes the first image in the query and the resulting kernel matrix in the query, the size of which is consistent with the feature map in the last layer. Fig. 11 shows the effect of the partial kernel matrix on the corresponding original image which is different among various pedestrians. The reason is that the kernel matrix of the attention mechanism is obtained by multiplying the actual position of the corresponding pixel on the feature map. The final highlight is more concentrated on certain parts.

5. Conclusion and future plans

In this paper, a novel end-to-end CNN framework for robust feature extraction of person re-ID is proposed. We use resnet-50 to extract basic deep features and obtain the weight for each feature point using the proposed attention model. The proposed attention mechanism assigns higher weighting values to indicate spatial relationship between local features and global features, and use different nonlinear functions to carry out effective feature combination. To extract similar characteristics within individuals and differences between different persons, we combine two loss functions to obtain more representative and robust features for re-identifying persons in challenging scenes. Experiments show that our approach outperforms state-of-the-art results.

Adopting feature maps that fuse different layers would improve the performance of the system. However, the number of kernel matrix parameters gets higher since the feature map size of lower layers is large. Considering region information to

Fig. 11. Visualization of the self-attention model on the Market-1501. Our approach highlights distinctive image regions which are useful for person re-identification. The attention models primarily focus on foreground regions and generally correspond to specific body parts.

improve the computational efficiency when evaluating the image saliency is a promising direction.

Acknowledgments

The work is supported by the following projects: National Natural Science Foundation of China, Nr.: 61702322, 6177051715; The Essential project of Shanghai Science and Technology Committee, China Nr.: 18511101600.

References

[1] Y. Zhang, Y. Yang, T. Li, H. Fujita, A multitask multiview clustering algorithm in heterogeneous situations based on LLE and LE, Knowl.-Based Syst. 163 (2019) 776–786.
[2] D. Cheng, Y. Gong, S. Zhou, Person re-identification by multi-channel parts-based CNN with improved triplet loss function, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1335–1344, http://dx.doi.org/10.1109/CVPR.2016.149.
[3] Z. Zheng, L. Zheng, Y. Yang, Pedestrian alignment network for large-scale person re-identification, IEEE Trans. Circuits Syst. Video Technol. (2018) http://dx.doi.org/10.1109/TCSVT.2018.2873599.
[4] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 770–778.
[5] M. Geng, Y. Wang, T. Xiang, Y. Tian, Deep transfer learning for person reidentification, IEEE Trans. Image Process. (2016) 5576–5588.
[6] Y. Lin, L. Zheng, Z. Zheng, Y. Wu, Y. Yang, Improving person re-identification by attribute and identity learning in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp. 1–12.
[7] J. Bi, C. Zhang, An empirical comparison on state-of-the-art multi-class imbalance learning algorithms and a new diversified ensemble learning scheme, Knowl.-Based Syst. 158 (2018) 81–93.
[8] R. Varior, M. Haloi, G. Wang, Gated siamese convolutional neural network architecture for human re-identification, in: European Conference on Computer Vision(ECCV), Vol. 9912, 2016, pp. 791–808.
[9] 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 (2017) 3492–3506.
[10] 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 (CVPR), 2017, pp. 1320–1329.
[11] A. Hermans, L. Beyer, B. Leibe, In defense of the triplet loss for person reidentification, 2017, pp. 1–17, arXiv:1703.07737.
[12] 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(ECCV), Vol. 9911, 2016, pp. 135–153.
[13] L. Zheng, Y. Huang, H. Lu, Y. Yang, Pose invariant embedding for deep person reidentification, IEEE Trans. Image Process. (2017) 1–10.
[14] 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 (CVPR), 2017, pp. 907–915.
[15] L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, Glad: Global-local-alignment descriptor for pedestrian retrieval, in: ACM Multimedia Conference(ACMMM), 2017, pp. 420–428.
[16] Z. Zheng, L. Zheng, Y. Yang, Unlabeled samples generated by gan improve the person re-identification baseline in vitro, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3774–3782, http://dx.doi.org/10.1109/ICCV.2017.405.
[17] Z. Zhong, L. Zheng, Z. Zheng, Camera style adaptation for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 5157–5166, http://dx.doi.org/10.1109/CVPR.2018.00541.
[18] L. Wei, S. Zhang, W. Gao, Q. Tian, Person Transfer GAN to Bridge Domain Gap for Person Re-Identification, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1–10.
[19] X. Qian, Y. Fu, W. Wang, T. Xiang, J. Qiu, Y. Wu, Y. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2018, pp. 661–678.
[20] L. Zhang, L. Zhang, B. Du, J. You, D. Tao, Hyperspectral image unsupervised classification by robust manifold matrix factorization, Inform. Sci. 485 (2019) 154–169, http://dx.doi.org/10.1016/j.ins.2019.02.008.
[21] H. Yu, W. Zheng, A. Wu, X. Guo, S. Gong, J. Lai, Unsupervised person re-identification by soft multilabel learning, in: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)(2019), 2019.
[22] H. Wang, Y. Yang, B. Liu, H. Fujita, A study of graph-based system for multi-view clustering, Knowl.-Based Syst. 163 (2019) 1009–1019.
[23] P. Peng, T. Xiang, Y. Wang, M. Pontil, Unsupervised cross-dataset transfer learning for person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2016, pp. 1306–1315.
[24] Y. Chen, X. Zhu, S. Gong, Deep association learning for unsupervised video person re-identification, in: British Machine Vision Conference, (BMVC), 2018, pp. 1–10.
[25] J. Lv, W. Chen, Q. Li, C. Yang, Unsupervised Cross-dataset Person re-identification by transfer learning of spatial-temporal patterns, in: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2018, pp. 7948–7956.
[26] G. Song, B. Leng, Y. Liu, T. He, R. Cong, S. Cai, Region-based Quality Estimation Network for Large-scale Person Re-identification, The Association for the Advance of Artificial Intelligence(AAAI), 2017, pp. 1–8.
[27] W. Li, X. Zhu, S. Gong, Harmonious Attention Network for Person Re-Identification, in:IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), 2018, pp. 1–10.
[28] S. Li, S. Bak, P. Carr, X. Wang, Diversity Regularized Spatiotemporal Attention for Video-based Person Re-identification, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, (CVPR), 2018, pp. 1–12.
[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. (2015) 211–252.
[30] H. Luo, Y. Gu, X. Liao, S. Lai, W. Jiang, Bag of tricks and A strong baseline for deep person re-identification, in: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR)(2019), 2019.
[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, I. Polosukhin, Attention Is all you need, in: Conference on Neural Information Processing Syetems (NIPS), 2017, pp. 1–15.
[32] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the Inception Architecture for Computer Vision, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), 2016.
[33] 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(ICCV), 2015, pp. 1116–1124.

[34] 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(ECCV), 2016, pp. 17–35.
[35] 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 (CVPR), 2014, pp. 152–159.
[36] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (2010) 1627–1645.
[37] Y. Zhang, T. Xiang, T. Hospedales, H. Lu, Deep Mutual learning, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp.1–10.
[38] W. Li, X. Zhu, S. Gong, Person re-identification by deep joint learning of multi-loss classification, in: International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 2194–2200.
[39] 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 (CVPR), 2017, pp. 3376–3385.
[40] Z. Zhong, L. Zheng, D. Cao, S. Li, Re-ranking person re-identification with k-reciprocal encoding, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 3652–3661, http://dx.doi.org/10.1109/CVPR.2017.389.
[41] C. Su, J. Li, S. Zhang, J. Xing, W. Gao, Q. Tian, Pose-driven deep convolutional model for person re-identification, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3980–3989, http://dx.doi.org/10.1109/ICCV.2017.427.
[42] D. Chen, Z. Yuan, B. Chen, N. Zheng, Similarity learning with spatial constraints for person re-identification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1268–1277, http://dx.doi.org/10.1109/CVPR.2016.142.
[43] L. Zhang, T. Gong, Learning a discriminative null space for person re-identification, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1239–1248, http://dx.doi.org/10.1109/CVPR.2016.139.
[44] Z. Zheng, L. Zheng, Y. Yang, A discriminatively learned cnn embedding for person re-identification, ACM Trans. Multimed. Comput. Commun. Appl. 14 (2016) 1–10, http://dx.doi.org/10.1145/3159171.
[45] S. Bai, X. Bai, Q. Tian, Scalable person re-identification on supervised smoothed manifold, in: IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017, pp. 3356–3365.
[46] A. Schumann, R. Stiefelhagen, Person re-identification by deep learning attribute-complementary information, in: Computer Vision and Pattern Recongnition Workshops(CVPRW), 2017, pp. 1435–1443.
[47] Y. Sun, L. Zheng, W. Deng, S. Wang, Svdnet for pedestrian retrieval, in: IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3820–3828.
[48] Y. Chen, X. Zhu, S. Gong, Person re-Identification by deep learning multi-scale representations, in: 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, pp. 2590–2600, http://dx.doi.org/10.1109/ICCVW.2017.304.