Multimedia Tools and Applications (2022) 81:4649–4667
https://doi.org/10.1007/s11042-020-10494-4

1193: INTELLIGENT PROCESSING OF MULTIMEDIA SIGNALS

Self-attention mechanism in person re-identification models

Wenbai Chen[1] · Yue Lu[1] · Hang Ma[1] · Qili Chen[1] · Xibao Wu[1] · Peiliang Wu[2,3]

Received: 8 October 2020 / Revised: 26 November 2020 / Accepted: 29 December 2020 /
Published online: 17 February 2021

Abstract

In recent years, person re-identification based on video has become a hot topic in the field of person re-identification. The self-attention mechanism can improve the ability of deep neural networks in computer vision tasks such as image classification, image segmentation and natural language processing tasks. In order to verify whether the self-attention can improve the performance or not in person re-identification tasks, this paper applies two self-attention mechanisms, non-local attention and recurrent criss-cross attention to person re-identification model, and experiments are conducted on Market-1501, DukeMTMC-reID and MSMT17 person re-identification datasets. The results show that the self-attention mechanism can improve the accuracy of the person re-identification model. The accuracy is higher when the self-attention module is inserted into the convolutional layers of the re-identification network.

✉ Wenbai Chen
chenwb@bistu.edu.cn

Yue Lu
luyue163@126.com

Hang Ma
mh@bistu.edu.cn

Qili Chen
qilichen@hotmail.com

Xibao Wu
wuxibao@bistu.edu.cn

Peiliang Wu
peiliangwu@ysu.edu.cn

1 School of Automation, Beijing Information Science & Technology University, Beijing 100192, China

2 School of Information and Engineering, Yanshan University, Qinhuangdao 066004, China

3 State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

Keywords Person re-identification · Deep neural network · Self-attention · Computer vision

1 Introduction

Person re-identification is widely considered as a subproblem of image retrieval [18, 24, 25], which uses computer vision technology to determine whether there is a specific person in the image or video. That is, given a monitored person image, the person image under cross equipment is retrieved [5, 27, 29]. Person re-identification technology can make up for the visual limitations of the current fixed camera and can be combined with person detection and person tracking technology, which can be used in video monitoring, intelligent security and other fields.

In recent years, deep learning represented by the convolutional neural network has achieved great success in the field of computer vision. It has defeated traditional methods in many tasks and even surpassed the level of human beings to some extent. In the person re-identification problem, the method based on deep learning can automatically learn complex feature descriptions, and using simple Euclidean distance to measure the similarity can achieve good performance [24]. At present, the person re-identification method based on deep learning has greatly surpassed the traditional method in performance. These advantages make deep learning popular in the field of person re-identification, and the research of person re-identification has entered a new stage [22].

The typical process of person re-identification is shown in Fig. 1. For the camera and captured images/videos, person detection is carried out first to obtain person images [24]. In order to eliminate the effect of person detection on the re-identification results, most person re-identification algorithms use the cropped person image as input. Then, stable and robust features are extracted from the input image to obtain the feature expression vectors which can describe and distinguish the different person. Finally, the similarity measurement is carried out according to the feature expression vector, and the images are sorted according to the similarity. The image with the highest similarity will be regarded as the final recognition result. Person re-identification includes two core parts: 1) feature extraction and expression. From the appearance of the person, the feature representation vectors with strong robustness and strong discrimination are extracted to effectively express the characteristics of the person images; 2) similarity measurement. The similarity of person is judged by the similarity comparison between feature vectors. It can be seen that the idea of person

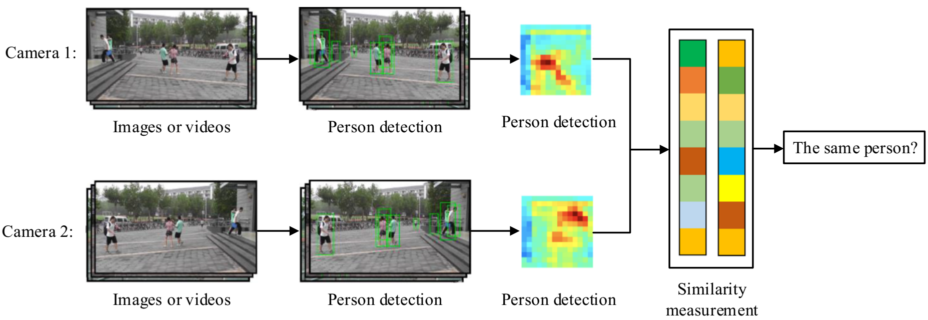

Fig. 1 Typical process of person re-identification

re-identification is the same as that of image retrieval, which can be regarded as a subproblem of image retrieval. According to the data source of person re-identification, it can be divided into image-based and video-based person re-identification. The latter benefits from more abundant time information in the video [15], which can obtain better performance.

The picture of the person under the camera is shown in Fig. 2. It can be seen that the research of person re-identification faces many challenges, such as low image resolution, angle change, posture change, light change and occlusion [5]. 1) In general, the face of surveillance video is fuzzy and the resolution is low. Therefore, face recognition and other methods cannot be used for re-identification. Only the human appearance information outside the head can be used for recognition. However, the body shape and clothing of different people may be the same, which brings great challenges to the accuracy of person re-identification. 2) The images of person re-identification are often collected from different cameras, due to the different shooting scenes and camera parameters, person re-identification generally has problems such as illumination change and angle of view change, which leads to the great difference of the same person under different cameras, and the appearance characteristics of different person may be more similar than that of the same person. 3) The person images for re-identification may be shot at different times. The posture and clothes of the person will change in different degrees. Besides, the appearance

Fig. 2 person pictures taken by the camera

characteristics of person will vary greatly under different lighting conditions. In addition, the scene under the actual video monitoring is very complex. Many monitoring scenes have a large flow of people and complex scenes. It is easy to block the beautiful face. In this case, it is difficult to re-recognize by gait and other features. The above situations have brought great challenges to the research of person re-identification, so the current research is still far from the practical application level.

Attention mechanism [19, 20] is a method to map a query and key pair to the output. The output is obtained by a weighted sum of weights, which is the similarity between query and key. Self-attention belongs to a kind of attention mechanism. Self-attention only acts on an object. It is widely used in machine translation and computer vision. The achievements of attention mechanisms in natural language processing have attracted worldwide attention. The natural language processing model based on the Transformer has made outstanding achievements in many aspects. Attention mechanism has also attracted extensive attention and research in computer vision, especially non-local attention networks. With the development of the network, the self-attention mechanism is combined with a convolutional neural network skillfully, which makes many models improve in various computer vision tasks. Literature [13] proposes a novel attention convolutional neural network, which is lightweight and can learn global and local features jointly. In [30], a from-top-to-bottom attention mechanism network is constructed, to enhance the saliency of the spatial pixel feature, showing improvements in re-identification tasks. The non-local module is a classical self-attention module in the computer vision field and has strong global feature extraction capability. Therefore, we adopt the non-local module as one of the self-attention modules for experiments. However, the non-local module also has some shortcomings. For example, it takes up a lot of memory for large feature maps. Some researchers have proposed some improvement methods, such as recurrent criss-cross attention (RCCA) [8] network transforms one-time global operation into two cross path operations, which can effectively reduce the memory consumption of large feature map. Thus we also use the RCCA module to apply self-attention to re-identification networks.

To verify whether the self-attention is beneficial for person re-identification models, we first add a non-local module in the middle of the baseline model to verify whether it can improve the re-identification model; then replace non-local with RCCA module to verify whether the RCCA module is effective; because the final extracted features have a direct impact on the similarity, we try to use the classifier normalization layer before adding 1D self-attention module, to verify whether it will bring improvement.

2 Related work

2.1 Person re-identification

Before the person re-identification technology was formally proposed, person re-identification was closely connected with multi-camera tracking technology [22]. In 1997, reference [7] proposed using the Bayesian formula to estimate the posterior probability of the appearance of an object in a camera under the premise of observing other camera views. The appearance model includes color, vehicle length, height and width, speed and observation time. The technical concept of "person re-identification" was first proposed in the 2006 document [3], which assumes that each person has a person tag. Bayesian network is used to predict the probability relationship between the person tag and the person characteristics, to predict the person's identity.

In 2006, literature [3] proposed that only human visual cues can be used for foreground detection through spatiotemporal segmentation algorithm. Visual matching is based on color and edge histogram, and is completed by a joint person model or Hessian affine point of interest operator. The experiment was carried out on a dataset. A total of 44 people were captured by three cameras and had a certain degree of visual occlusion. Reference [3] designed a spatiotemporal segmentation method, but did not use spatiotemporal motion information, which can be classified as image-based person re-identification.

For the video tracking task, most of person re-identification work focuses on image matching [27]. In 2010, two multiple re-identification tasks were proposed in [1] and [2], which were based on video. Both of them use color features. In addition, the segmentation model is used to detect the foreground in reference. For distance measurement, the minimum distance between the bounding boxes in two image sets is calculated in both works. The Bhattacharyya distance is further used for color and general features in reference [1]. The results show that using multiple frames can improve re-identification accuracy.

The success of deep learning in image classification [11] was extended to person re-identification in 2014, and literature both [23] and [12] used siamese neural networks to determine whether a pair of input images belong to the same identity. The reason why we choose the twin network is that the number of each identity sample is very small. The main difference between the two is that several cost loss functions have been added in reference [23], while reference [12] has made a precise regional division of the person body. Reference [23] and [12] do not use the same dataset, so the two methods cannot be directly compared.

At present, the person re-identification method based on deep learning has greatly surpassed the traditional method in performance. Since the generated countermeasure network [4] can effectively expand the dataset and make up for the lack of person re-identification data, there are many person re-identification studies based on the generated countermeasure network; besides, semi-supervised learning, unsupervised learning, and transfer learning are also important research directions [5]. In recent years, people pay more and more attention to the end-to-end deep learning methods, such as the representative siamese convolutional neural network (SCNN) [23], filter pairing neural network (FPNN) [12], and the recent excellent methods are harmonic attention network [13], joint critical and generative learning for person re-identification [28], etc. The emergence of large datasets also provides strong support for person re-identification methods based on deep learning. At present, the commonly used image datasets mainly include Market-1501 [26], DukeMTMC-reID [17], MSMT17 [21], CUHK03 [12], etc.

2.2 Residual neural network

In the neural network, gradient vanishing and gradient explosion are the problems easily encountered in the training process. Gradient explosion can be limited by gradient clipping, but gradient disappearance cannot be solved by intuitive restriction. The emergence of the residual network (ResNet) [6] effectively solves the problem of gradient disappearance. In the residual neural network, the input feature map passes through a convolution and activation layer, and then passes through a convolution and activation layer, and then it is added with the original feature map at the element level, and then passes through the second activation layer.

In ResNet, two different residual connection units are used in different networks. ResNet-50, ResNet-101 and ResNet-152 use the bottleneck structure. This kind of bottleneck unit first reduces the number of convolution layer channels through the convolution kernel of size 1, thus reducing the calculation amount of size 3 convolution layer, and then restores the original channel number through the convolution kernel of size 1, keeping the size of the feature map output before and after the residual unit unchanged. The combination of bottleneck and residual structure enables it to be applied to deep neural networks, but it does not bring too many parameters and calculations, and does not reduce the ability of feature extraction. From ResNet-34 to ResNet-50, their architectures are similar. Although the network depth has been increased by half, the amount of calculation has only increased a little. This is the advantage of using bottleneck, which can increase the network depth and improve the feature extraction ability of the network, while not increasing a lot of calculation. Because ResNet-50 has a more powerful feature extraction ability than ResNet-34, and the amount of calculation does not increase too much, so a lot of work will use ResNet-50 as a backbone network. Based on the above conditions, ResNet-50 is also used as the baseline model.

3 Method

3.1 Self-attention module

Non-local (NL) [20] attention module is a classic application of the self-attention mechanism in computer vision. It makes use of a self-attention mechanism to make each position in the feature map interact with other positions, and get the correlation between two positions so that the network has a global view. The calculation formula of the non-local module of the embedded Gaussian method is shown in (1):

$$\begin{cases} y = softmax(\theta(x) \cdot \phi^T(x))g(x) \\ z = w(y) + x \end{cases} \tag{1}$$

in which x is the input tensor, θ, ϕ, g and w denote convolutional operations, and $softmax$ operation is defined as $softmax(x_i) = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}}$.

The schematic diagram of the embedded Gaussian method non-local module is shown in Fig. 3.

Firstly, query, key and value are generated by three convolution layers (θ, ϕ and g) with the size of 1 convolution kernel. Then, query and key are multiplied, and then the embedded Gaussian method is used to normalize ($softmax$) and multiply with value. Finally, a convolution layer w is used to obtain the feature graph with the same shape as the original feature graph, and the feature graph is made at the element level with input $\mathbf{X}$ added as the final output $\mathbf{Z}$. The number of channels of convolution layer θ and convolution layer ϕ can be smaller than the input characteristic graph $\mathbf{X}$, and can be set to 1/2, 1/4, 1/8 or 1/16 of $\mathbf{X}$. In this experiment, the number of channels of g can be arbitrary, and the number of output channels of w must be consistent with $\mathbf{X}$, to ensure that the output of the non-local module is the same as the size of input characteristic graph, and can be inserted into any position without changing the network structure layer. The output dimension of g is the same as that of θ and ϕ, which is 1/4 of the channel number of $\mathbf{X}$, which can effectively reduce the amount of calculation and memory usage. The 2D non-local module used in this experiment is in this form. In addition, the non-local module can also downsample the feature maps of ϕ and g to further reduce the size of the feature map, to reduce the amount of calculation and

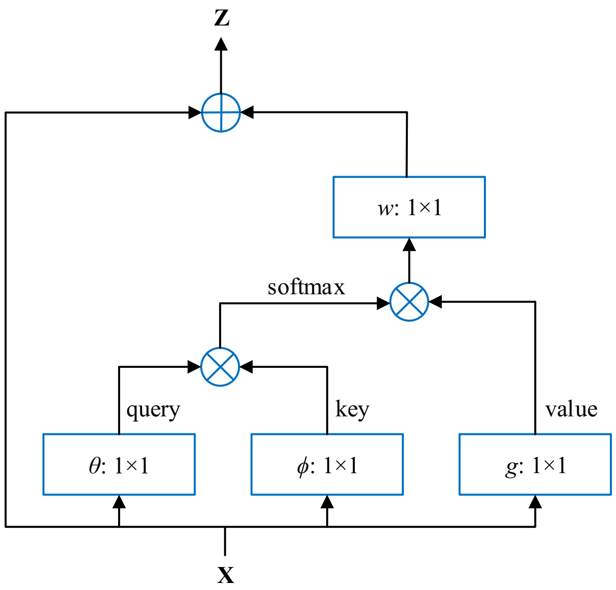

Fig. 3 Non-local schematic diagram

memory usage. But in this experiment, the original non-local calculation method is used instead of downsampling the feature map.

The non-local module can enhance the ability of global feature extraction. Because of the same size of input and output feature graph, it can be easily inserted into any place of network structure, and has strong applicability. In order to solve the problem that non-local takes up too much memory when the feature map is too large, a recurrent criss-cross attention (RCCA) module is proposed to reduce the complexity of the attention module to

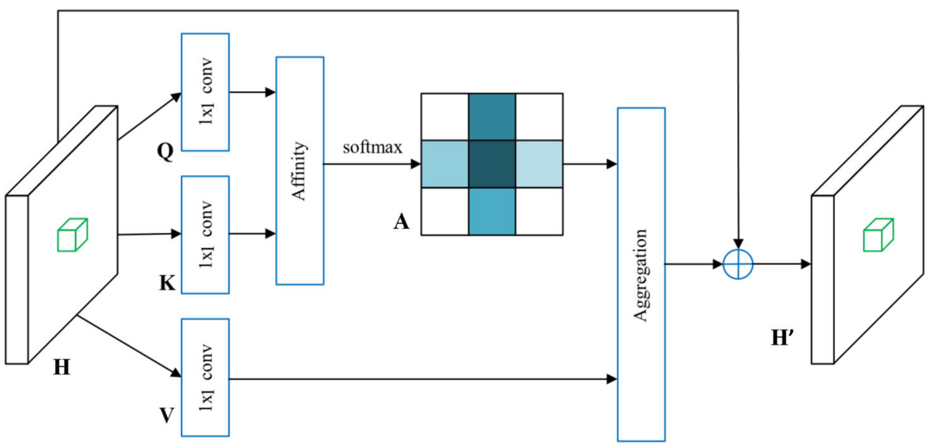

Fig. 4 Schematic diagram of the criss-cross attention module

reduce the memory consumption. The schematic diagram of a CCA module is shown in Fig. 4. Affinity and Aggregation are the keys to criss-cross attention [8].

At each position u in spatial dimension of feature maps $\mathbf{Q}$, we can get a vector $\mathbf{Q}_u \in \mathbb{R}^{C'}$, where C' is the number of channels in $\mathbf{Q}$ and $\mathbf{K}$. Meanwhile, we can obtain the set $\Omega_{\mathbf{u}}$ by extracting feature vectors from $\mathbf{K}$ which are in the same row or column with position u. Thus, $\Omega_{\mathbf{u}} \in \mathbb{R}^{(H+W-1)\times C'}$. $\Omega_{\mathbf{i,u}} \in \mathbb{R}^{C'}$ is the ith element of $\Omega_{\mathbf{u}}$. The Affinity operation is defined as follows:

$$d_{i,u} = \mathbf{Q_u}\Omega_{\mathbf{i,u}}^{\mathsf{T}} \tag{2}$$

in which $d_{i,u} \in \mathbf{D}$ denotes the degree of correlation between feature $\mathbf{Q_u}$ and $\Omega_{\mathbf{i,u}}$, $i = [1, ..., |\Omega_{\mathbf{u}}|]$, $D \in \mathbb{R}^{(H+W-1)\times W\times H}$. Then, we apply a $softmax$ layer on $\mathbf{D}$ along the channel dimension to calculate the attention map $\mathbf{A}$.

Then another convolutional layer with 1×1 filters is applied on $\mathbf{H}$ to generate $\mathbf{V} \in \mathbb{R}^{C\times W\times H}$ for feature adaption. At each position u in spatial dimension of feature maps $\mathbf{V}$, we can obtain a vector $\mathbf{V_u} \in \mathbb{R}^{C}$ and a set $\Phi_{\mathbf{u}} \in \mathbb{R}^{(H+W-1)\times C}$. The set $\Phi_{\mathbf{u}}$ is collection of feature vectors in $\mathbf{V}$ which are in the same row or column with position u. The long-range contextual information is collected by the Aggregation operation:

$$\mathbf{H}'_{\mathbf{u}} = \sum_{\mathbf{i}\in|\Phi_{\mathbf{u}}|} \mathbf{A_{i,u}}\Phi_{\mathbf{i,u}} + \mathbf{H_u} \tag{3}$$

in which $\mathbf{H}'_{\mathbf{u}}$ denotes a feature vector in output feature maps $\mathbf{H}' \in \mathbb{R}^{C\times W\times H}$ at position u. $\mathbf{A_{i,u}}$ is a scalar value at channel i and position u in $\mathbf{A}$. The computational complexity of the non-local module is $O((H \times W) \times (H \times W))$, while that of the RCCA module is $O((H \times W) \times (H + W - 1))$, which is smaller than the non-local module.

3.2 Baseline model

In this paper, ResNet-50 is selected as the benchmark network and backbone. Based on the original ResNet-50, a batch normalization [9] layer is added before the final classifier to improve the robustness of the network and improve the re-identification accuracy. Because the person re-identification task is different from the image classification task with a fixed number of categories, the whole connection layer is not immutable. In training, the output dimension of the classifier is set to be consistent with the number of personal identities in the training set. During the test, because the number of personal identities changes and is generally not associated with the person identity during training, the classifier needs to be removed and only the feature vector obtained before the classifier is retained. Then, according to the similarity between the feature vector and the feature vector extracted from the person images in the retrieval database, the classifier needs to be removed. Further, the model was evaluated by the evaluation criteria.

In human re-identification, it is usually necessary to classify the person's identity, which can be met by using cross-entropy loss. Besides, it is necessary to increase the distance between the images of the person with different identities and reduce the distance between images of the person with the same identity, and triple loss can be used.

Usually, the $softmax$ layer is used to normalize the output vector to 0 to 1, and the sum of all elements of the vector is 1, and then a cross-entropy loss is used. In order to simplify the calculation, $softmax$ and cross-entropy loss are usually combined, and the calculation process is shown in (4).

$$loss(x, class) = -log\left(\frac{exp(x[class])}{\sum_j exp(x[j])}\right) = -x[class] + log\left(\sum_j exp(x[j])\right) \tag{4}$$

where x means that the forward propagation in the training process of the network model does not pass the output of $softmax$. The two steps of $softmax$ processing and calculating cross-entropy loss are included in formula (4), which is the label corresponding to the input image, that is, the category it should be classified, where is the total number of categories and the total number of identities corresponding to the training set.

The input of triplet loss is three elements, which are a real element, a similar element and a different class element. Through triple loss, the loss between the real element and similar element can be reduced, and the loss between different type element can be increased so that the same type element is more similar and the different class element is more different. The calculation process is shown in (5).

$$loss(x^a, x^p, x^n) = \sum_i^N \left[dist(x_i^a, x_i^p) - dist(x_i^a, x_i^n) + \alpha\right]_+ \tag{5}$$

In the above formula, x^a, x^p, x^n respectively represents the anchor, positive and negative elements, and $dist$ represents the distance function. It can be Euclidean distance or cosine similarity. It is a distance interval parameter. It is used to set the distance between similar elements and real elements and the distance gap between different class elements and real elements. The $+$ outside [] means that the value in [] is greater than or equal to 0, and the value in [] is set to 0 when it is less than 0.

3.3 Applying attention modules

In Section 3.1, we mentioned that non-local modules can be inserted into any stages in a network because the input and output of a non-local module have the same shape. For example, to insert a non-local module into the conv3_1 layer of ResNet-50, we need to feed the features from the conv3_1 layer to the non-local module. The non-local module outputs the features calculated by (1). Then we feed the output features from the non-local module to the next layer conv3_2. As a result, The ResNet-50 network is equipped with one non-local module. A network can be inserted into multiple non-local modules at different locations simultaneously. We give the algorithm pseudo codes of the proposed method as shown in Algorithm 1. The application of RCCA modules is the same as non-local modules. The architecture of the whole model is shown as Fig. 5.

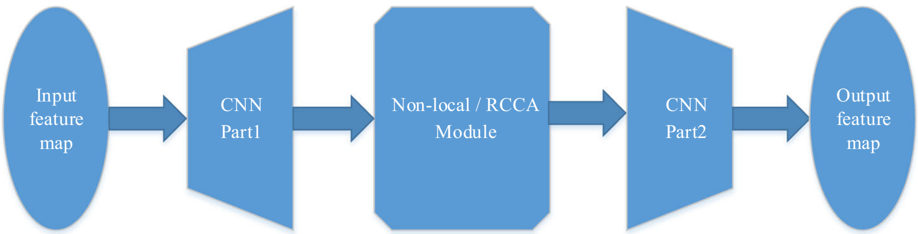

Fig. 5 The architecture of the model

Algorithm 1 Applying Non-local module.

Require: $Network$:= the CNN model
Require: $Data$:= the data to be feed into the model
Require: $LayerList$:= the list of the convolutional layers to be inserted into the non-local module

function NON_LOCAL_MODULE($feature\_map$)
    $feature\_map$:= the input feature map which is a tensor
    $X \leftarrow feature\_map$
    θ, ϕ, g, w := convolutional operations
    $Q \leftarrow \theta(X)$
    $K \leftarrow \phi(X)$
    $V \leftarrow g(X)$
    $F_1 \leftarrow Q \cdot K$
    $F_2 \leftarrow softmax(F_1)$
    $Y \leftarrow F_2 \cdot V$
    $Z \leftarrow w(Y) + X$
    return Z
end function
function INSERT_NL($model, input, layers$)
    $model$:= the CNN model
    $input$:= the input data
    $layers$:= the the layers to be inserted into the non-local module
    $X \leftarrow input$
    for each $layer_i \in models$ **do**
        $X \leftarrow layer_i(X)$
        if $conv_i \in layers$ **then**
            $X \leftarrow$ NON_LOCAL_MODULE(X)
        end if
    end for
    $Y \leftarrow X$
    return Y
end function
$OutputTensor \leftarrow$ INSERT_NL($Network, Data, LayerList$)

4 Experiment

4.1 Dataset

At present, there are many image datasets for person re-identification. Three larger datasets are used in this paper, namely Market-1501, DukeMTMC-reID and MSMT17. The details of each dataset are shown in Table 1.

4.2 Data augmentation

After reading the image data, it is necessary to preprocess the image data to achieve the function of data expansion, which is conducive to reducing overfitting. However, the method of data preprocessing in training is different from that in testing. During training, the image is

Table 1 Comparison of three datasets

Dataset	Release	Identities	Images	Cameras	Marking method	Resolution	Scene
Market-1501	2015	1501	32688	6	DPM	fixed	outdoor
DuckMTMC	2017	1404	36411	8	manual	fixed	outdoor
MSMT17	2018	4101	126441	15	Faster RCNN [16]	change	Outdoor and indoor

first scaled and randomly cropped, and then flipped at a random level with a probability of 0.5, which can be converted into a tensor. Then, the pixel value of the input image is normalized by using the normalization parameter. Finally, a small area is randomly selected in the image to erase the pixels in the region (make it 0). There are some differences between the preprocessed data and the original data, and each picture will have a different form, which is equivalent to getting several similar images, which plays the role of data expansion. During the test, fixed scaling is performed first, then converted into tensor, and finally normalized, which ensures that each test image is the same after processing. In order to make full use of the image information, the features of the image itself and the features of the image after horizontal flipping are taken at the same time during feature extraction, and the two features are added as the overall features of the image.

4.3 Criterion

In this paper, two commonly used evaluation criteria for person re-identification tasks are used: top-1 accuracy and mean average precision (mAP). Top1 accuracy is the accuracy of the first retrieval object to be retrieved, but there will be multiple real images to be retrieved in person re-identification task, and mAP can measure the average retrieval performance of multiple real images. At first, the mAP is widely used in image retrieval, and it also can solve the same problem in two aspects of person image retrieval.

4.4 Experimental setup

All experiments adopt the same settings. All experimental settings are shown in Table 2. We adopt the learning rate attenuation strategy. The initial learning rate is set to 0.00035. Every

Table 2 Parameter settings

Parameter	Value
Learning rate	0.00035
Batch	size 48
Epoch	100
Learning rate decay	30,60,90
Gamma	0.1
Height, width	(256,128)
Optimizer	Adam [10]
Weight decay	5e-04
Loss	cross entropy loss+triplet loss
Margin	0.3
Distance	cosine

Table 3 Results of the baseline

Dataset	Top1 accuracy	mAP
Market-1501	94.5%	85.9%
DukeMTMC-reID	84.4%	73.5%
MSMT17	70.3%	45.4%

30 rounds, the learning rate is multiplied by 0.1, and a total of 100 rounds of training are conducted. In the triple loss function, we need to set a boundary value to judge whether to optimize. We set the boundary value to 0.3, and the cosine distance is used as the distance function. In addition, the ResNet-50 model is initialized with the parameters trained on ImageNet [11].

The GPU we use is an NVIDIA Titan XP with 12GB memory. The CPU is E5-2678, and we only use 5 cores. We limit the RAM usage to 10 GB, which is enough for all the experiments. Our operating system is Ubuntu-16. All the experiments are constructed with PyTorch-1.2 [14] framework.

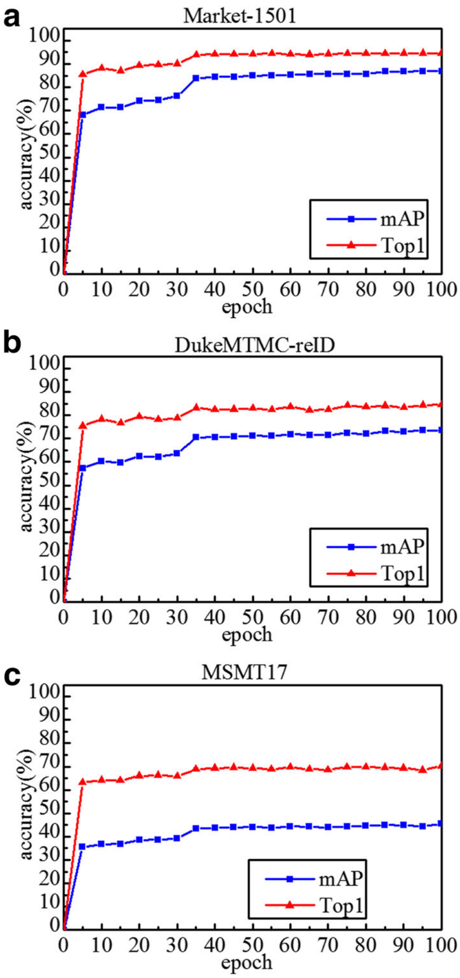

Fig. 6 Trend of the accuracy on the three datasets. **a** Market-1501. **b** DukeMTMC-reID. **c** MSMT17

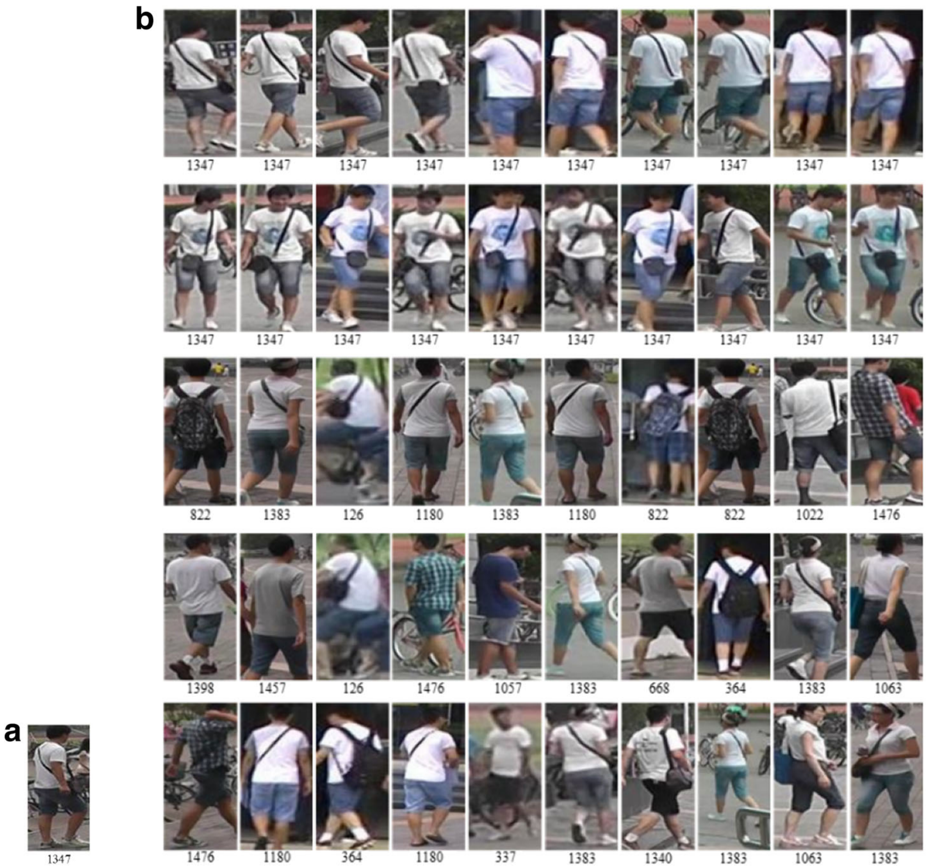

Fig. 7 Visualization of the results. **a** Query. **b** Top 50 searched results

4.5 Baseline results

The results of the ResNet-50 baseline model on three datasets are shown in Table 3. During the training process, we test every 5 rounds. The changing trend of test accuracy is shown in Fig. 6. The red curve represents the Top1 accuracy rate, and the blue curve represents the mAP. It can be seen that due to the reduction of the learning rate in the 30th round, the accuracy rate suddenly increased, indicating that the learning rate before the 30th round has optimized the model to near saturation; in the later stage, the learning rate changes slowly

Table 4 Results of baseline adding non-local module

Dataset	Top1 accuracy	mAP
Market-1501	95.2%(+0.7%)	87.4%(+1.5%)
DukeMTMC-reID	85.2%(+0.8%)	75.1%(+1.6%)
MSMT17	71.0%(+0.7%)	46.8%(+1.4%)
Average increment	0.73%	1.5%

with a little jitter, but it tends to be stable overall. The difficulty of the three datasets is different. The Market-1501 is easy to get a relatively high accuracy rate, and MSMT17 is more complex, so the accuracy rate is relatively low.

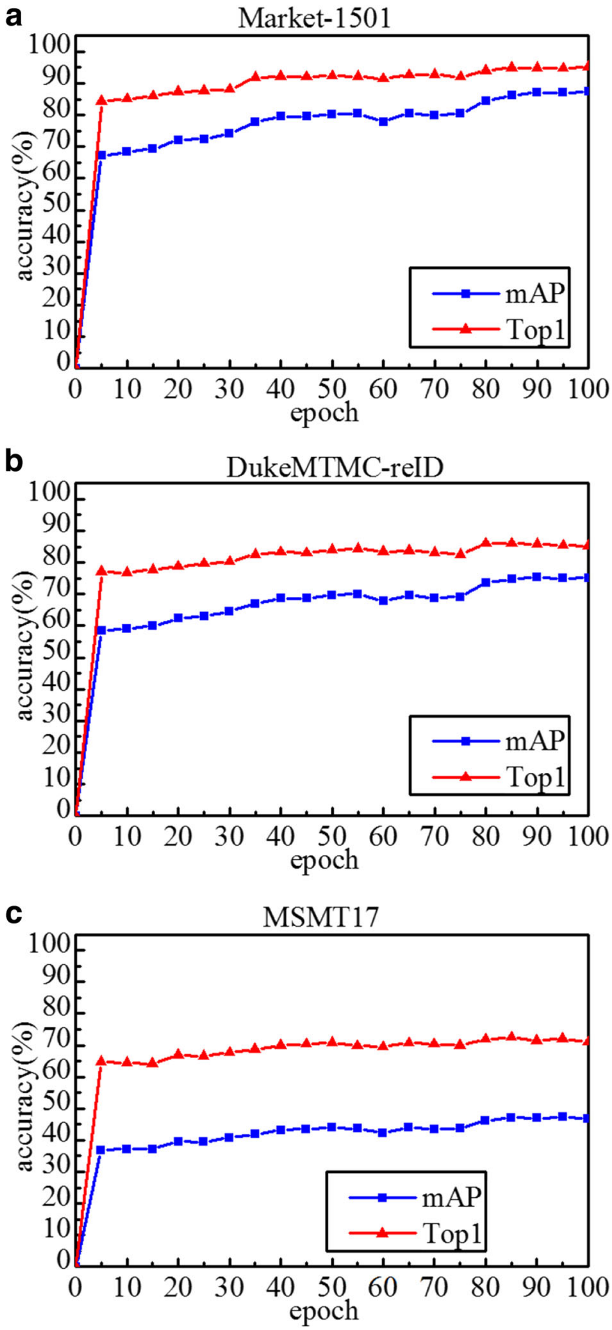

Fig. 8 Trend of the accuracy on the three datasets. **a** Market-1501. **b** DukeMTMC-reID. **c** MSMT17

4.6 Visualization of results

In order to intuitively understand the reasoning results and practical application effect of the model, it is necessary to visualize the results of the model. Take the image in the Market-1501 dataset as an example, select a query image, as shown in Fig. 7a, whose person ID is 1347. Through the trained baseline model with pre-training parameters, the top 50 most similar person images retrieved in the retrieval database are displayed, as shown in Fig. 7b. The number below each person's image represents the person's identity.

From the visualization results, the person ID of the first 20 retrieved images is 1347, while there are 20 images with the ID of 1347 taken by different cameras from the query images in the retrieval database, which means that the first 20 images are completely correct and all the correct images have been found. In the first 50 retrieved images, most of the person in the first 50 retrieval images have shoulder bags and wear light-colored coats and blue or gray bottoms, which also shows that the model can capture color and texture information well and distinguish a different person.

4.7 Non-local module added to convolution layer

Conv3_1, conv3_3, conv4_1, conv4_3, conv4_5 in ResNet-50 are inserted with the non-local 2D module respectively, and test it on Market-1501 to verify whether adding a non-local can improve the performance of the model. The experimental results of adding non-local module to the baseline are shown in Table 4. In the same training process, the test is conducted every 5 rounds, and the change trend curves of accuracy rate are shown in Fig. 8. Because of the use of pre-training, the overall convergence is faster, and after adding non-local, it can be seen that the test accuracy rate has been increasing in the later training process, which also causes the final accuracy rate to be higher than that without non-local; according to the final curve trend, continuing training may also increase the accuracy rate, but to keep the experimental conditions consistent, we still end the training.

The experimental results show that the addition of a non-local module can effectively improve the accuracy of ResNet-50 in person re-identification. The accuracy of top-1 and mAP are improved by 0.73% and 1.5% respectively, which can be said to be a great improvement.

4.8 RCCA module added to convolution layer

The RCCA 2D module is inserted in the same five positions, and the experiment is conducted on Market1501 to verify whether RCCA can improve the performance of the model. The experiment of adding the RCCA module to baseline is shown in Table 5. The curves of test accuracy on three datasets are shown in Fig. 9. The trend of adding the RCCA module is similar to that of non-local module. It converges to a better state very soon, but it is very

Table 5 Results of RCCA module added to baseline

Dataset	Top1 accuracy	mAP
Market-1501	95.1%(+0.6%)	87.7%(+1.8%)
DukeMTMC-reID	85.1%(+0.7%)	75.1%(+1.6%)
MSMT17	71.0%(+0.7%)	47.0%(+1.6%)
Average increment	0.67%	1.67%

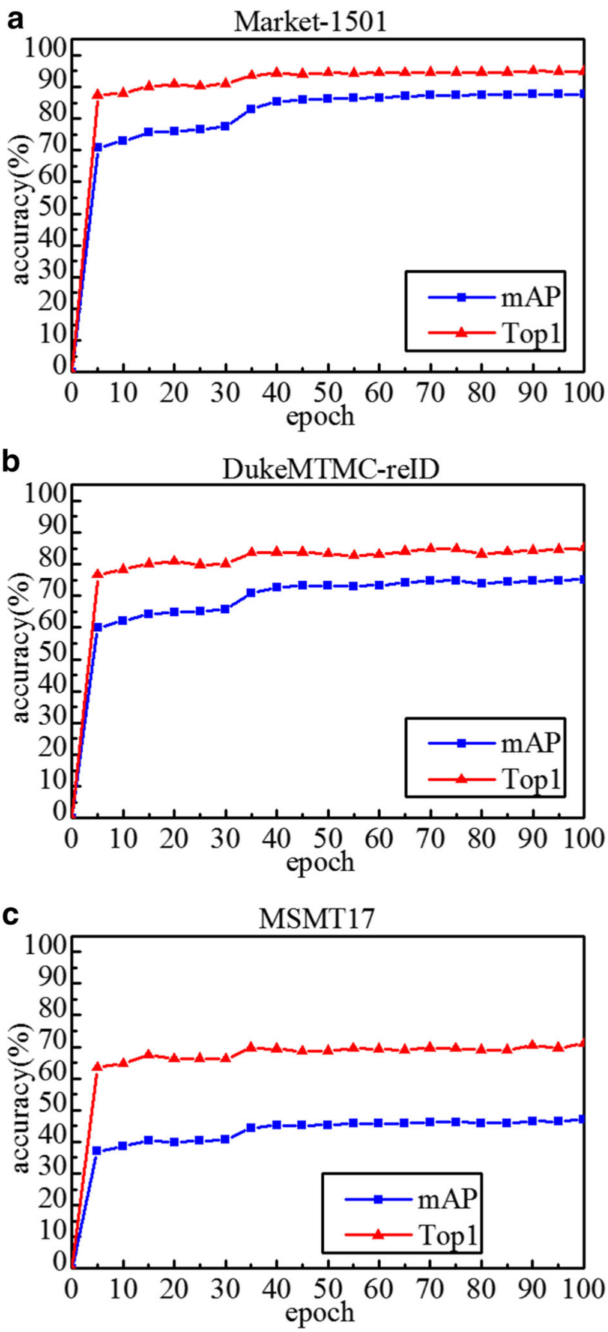

Fig. 9 Trend of the accuracy on the three datasets. **a** Market-1501. **b** DukeMTMC-reID. **c** MSMT17

Table 6 Adding non-local 1D module before BN layer

Dataset	Top1 accuracy	mAP
Market-1501	94.6%(+0.1%)	85.4%(-0.4%)

close to the final accuracy rate very early. The convergence speed of later training is faster than that of the non-local module.

It can be seen that the RCCA module and non-local module have similar functions, which can significantly improve the top-1 accuracy and mAP of person re-identification. The top-1 accuracy rate is increased by 0.67% on average, and the RCCA module has a stronger ability to improve the mAP, with an average increase of 1.67%.

4.9 Non-local module added before normalization layer

The non-local 1D module was added before the batch normalization (BN) layer of ResNet-50, and the experiment was conducted on Market-1501 to observe whether the accuracy could be improved. Non-local here. The module is constructed for 1D feature vector input, which is different from the module dimension for the 2D feature map. For the vector with an input dimension of 1D, the convolution layer should be changed to a 1D convolution layer, in which the matrix multiplication operation is also changed to 1D vector multiplication, and finally the output result is 1D. In this way, the non-local module can be inserted into a 1D feature map, while maintaining its output dimension and the dimensions of the original 1D vector are the same.

The non-local 1D module is added before the BN layer. The results are shown in Table 6, and the change curve of test accuracy is shown in Fig. 10. It can be seen that the training convergence of the non-local 1D module is faster, and there is a certain fluctuation from the early stage to the middle stage and from the middle stage to the later stage, which shows an upward trend as a whole and is relatively stable in the later stage.

In the experimental results, the accuracy of top-1 did not change much compared with the baseline, but the mAP decreased slightly. From the experimental results, this method can not make the re-identification model gain effectively, so no further experiments on other datasets are needed.

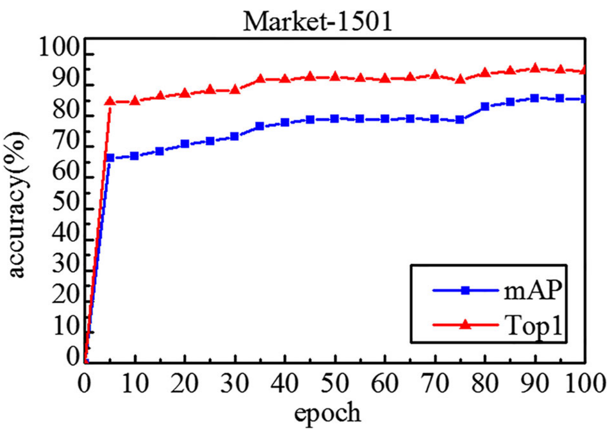

Fig. 10 Trend of the accuracy on the Market-1501

4.10 Empirical conclusion

According to the experimental results of adding non-local 2D, the mAP and the accuracy of top-1 can be effectively improved by adding non-local 2D to five conventional convolution layers. RCCA 2D module has the same effect as the non-local 2D module. According to the test trend chart, the convergence rate of the baseline added with the two methods is slower than that of the baseline without adding, but the final accuracy rate is higher than that of the baseline, which indicates that the non-local 2D or RCCA is added2D has some influence on the original pretraining parameters, which makes the accuracy rate appear to be lower than the baseline before the training is finished, but the final accuracy rate will be higher due to the role of the two modules. However, for the final output feature map of ResNet-50, adding a non-local 1D module can improve the performance of the re-identification model slightly, even slightly decrease, so it is not suitable to add an attention module before the BN layer.

5 Summary

In this paper, the attention mechanism is applied to person re-identification model, and two common attention modules, the non-local module and the RCCA module are explored. We implement a person re-identification model based on ResNet-50, and experiment on three datasets Market-1501, DukeMTMC-reID and MSMT17. We conclude that the performance of the person re-identification model can be improved when the attention mechanism is inserted into the convolution layer. However, when the size of the feature map is large, both of the two attention models still need a lot of computation and memory. Future work can be focused on the study of more lightweight attention mechanisms suitable for person re-identification tasks. In addition to considering the long-distance relationship between pixels in a picture, the pixel relationships between two images should also be considered when retrieving images. Therefore, how to use the attention mechanism to model the relationships between different pedestrians is also an aspect of future work.

Acknowledgments This work was partially supported by the National Key R&D Program of China (2018YFB1308300), China Postdoctoral Science Foundation (2018M631620), and partially by Cross-Training Plan of High Level Talents and Training Project of Beijing, and Beijing Natural Science Foundation(Grant No.4202026).

Compliance with Ethical Standards

Conflict of Interests The authors declare no conflict of interest

References

1. Bazzani L, Cristani M, Perina A, Farenzena M, Murino V (2010) Multiple-shot person re-identification by hpe signature. In: 2010 20th International conference on pattern recognition. IEEE, pp 1413–1416
2. Farenzena M, Bazzani L, Perina A, Murino V, Cristani M (2010) Person re-identification by symmetry-driven accumulation of local features. In: 2010 IEEE computer society conference on computer vision and pattern recognition. IEEE, pp 2360–2367
3. Gheissari N, Sebastian TB, Hartley R (2006) Person reidentification using spatiotemporal appearance. In: 2006 IEEE Computer society conference on computer vision and pattern recognition (CVPR'06), vol 2. IEEE, pp 1528–1535

4. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems, pp 2672–2680
5. Hao L, Wei J et al (2019) A survey on deep learning based person re-identification. ACTA AUTOMATICA SINICA 45(11):2032–2049. https://doi.org/10.16383/j.aas.c180154
6. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
7. Huang T, Russell S (1997) Object identification in a Bayesian context. In: IJCAI, vol 97, pp 1276–1282
8. Huang Z, Wang X, Huang L, Huang C, Wei Y, Liu W (2019) Ccnet: criss-cross attention for semantic segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 603–612
9. Jian-Wei L, Hui-Dan Z, Xiong-Lin L, Jun X (2020) Research progress on batch normalization of deep learning and its related algorithms. Acta Automatica Sinica 46(6):1090–1120
10. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. arXiv:1412.6980
11. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems, pp 1097–1105
12. Li W, Zhao R, Xiao T, Wang X (2014) Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 152–159
13. Li W, Zhu X, Gong S (2018) Harmonious attention network for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294
14. Paszke A, Gross S, Chintala S et al (2017) Automatic differentiation in pytorch
15. Qin L, Yu N, Zhao D (2018) Applying the convolutional neural network deep learning technology to behavioural recognition in intelligent video. Tehnicki Vjesnik 25:528–535. https://doi.org/10.17559/TV-20171229024444
16. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99
17. Ristani E, Solera F, Zou R, Cucchiara R, Tomasi C (2016) Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on computer vision. Springer, pp 17–35
18. SONG W, ZHAO Q et al (2017) Survey on pedestrian re-identification research. CAAI Trans Intell Syst 12(6):770–780. https://doi.org/10.11992/tis.201706084
19. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
20. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
21. Wei L, Zhang S, Gao W, Tian Q (2018) Person transfer gan to bridge domain gap for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 79–88
22. Ye M, Shen J, Lin G, Xiang T, Shao L, Hoi SC (2020) Deep learning for person re-identification: a survey and outlook. arXiv:2001.04193
23. Yi D, Lei Z, Liao S, Li SZ (2014) Deep metric learning for person re-identification. In: 2014 22nd International Conference on Pattern Recognition, pp. 34–39. IEEE
24. You-Jiao L, Li Z et al (2018) A survey of person re-identification. ACTA AUTOMATICA SINICA 44(9):1554–1568. https://doi.org/10.16383/j.aas.2018.c170505
25. Zajdel W, Zivkovic Z, Krose BJ (2005) Keeping track of humans: have i seen this person before? In: Proceedings of the 2005 IEEE international conference on robotics and automation. IEEE, pp 2081–2086
26. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q (2015) Scalable person re-identification: a benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
27. Zheng L, Yang Y, Hauptmann AG (2016) Person re-identification:, Past, present and future. arXiv:1610.02984
28. Zheng Z, Yang X, Yu Z, Zheng L, Yang Y, Kautz J (2019) Joint discriminative and generative learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2138–2147
29. Zheng Z, Zheng L, Yang Y (2017) A discriminatively learned cnn embedding for person reidentification. ACM Trans Multimed Comput Commun Applic (TOMM) 14(1):1–20
30. Ziyan L, Peipei W (2020) Pedestrian re-identification feature extraction method based on attention mechanism. J Comput Applic 40(3):672