Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-18875-9

Research on person re-identification based on multi-level attention model

Dan Wei[1] · Danyang Liang[1] · Longfei Wu[1] · Xiaolan Wang[1] · Lei Jiang[1] · Suyun Luo[1]

Received: 26 December 2023 / Revised: 25 February 2024 / Accepted: 9 March 2024

Abstract
Person Re-identification (ReID) is an important research direction in the field of pattern recognition, which aims to retrieve the same pedestrian in different cameras. The combination of deep learning and attention mechanism greatly improves the accuracy of image retrieval, but previous researchers usually use on-channel or spatial convolution to learn attention, ignoring the connection between attention feature nodes. In this article, we first improve a bottleneck attention module (BAM) to make the learned attention map faster. Secondly, to capture the relevance of each feature node in the global attentional feature map, we design a self-relevant attention module (SRA), which models the global scope structure information and is used to capture the connection between the feature node positions to make the obtained attentional map more robust. Finally, we propose a method to strengthen the attention features, so that the higher attention features around the position also get higher feature values, so that the obtained feature map is more robust. The effectiveness of the model is confirmed in several mainstream pedestrian re-identification datasets, and the proposed model outperforms most state-of-the-art methods.

Keywords Person re-identification · Bottleneck attention module · Self-relevant attention module · Feature extraction

1 Introduction

In recent years, with the rapid advancement of deep learning, it has led to a boom in many image fields. Such as remote sensing image[1], medical imaging[2–4] and etc. Surveillance video is widely used in various fields such as security, business, industrial production and intelligent transportation. According to statistics, about half of the world's hard disk is used for storing surveillance video, so it can be seen that surveillance video occupies a very important position in our daily life. The most important object of concern in the surveillance video is the pedestrians, identifying specific pedestrians for criminal investigation, violation of regulations has a very important significance.

Dan Wei
weiweidandan@163.com

[1] Shanghai University of Engineering Science, Shanghai, China

Published online: 16 March 2024

Person re-identification refers to matching specific pedestrians at different times, locations and cameras, which makes person re-identification a challenging topic due to its diverse pedestrian postures, confusing backgrounds and occlusions. In recent years, due to the development of deep learning, more and more researchers have started to focus on the research of person re-identification[5]. The ability of humans to focus their attention on a single recognizable feature in complex scenes has inspired researchers to introduce attention mechanisms into computer vision systems to improve recognition performance and reduce the negative effects of visual appearance, cluttered backgrounds, and pose changes[6].

Attention not only tells us where to focus, but also reinforces our areas of interest. Our goal is to increase representational ability by using attention mechanisms: focusing on important features and suppressing unnecessary ones. The SEnet network proposed by J. Hu et al. explicitly models the interdependencies between feature channels and does not introduce the spatial dimension for fusion between feature channels[7]. Specifically, it automatically obtains the importance of each feature channel by self-learning, and then boosts the useful features and suppresses the features that are not useful for the current task according to this importance. Woo et al. proposed the CBAM module[8], which sequentially infers the attention graph along two independent dimensions (channel and spatial) and then multiplies the attention graph with the input feature graph for adaptive feature optimization, and this author experimentally verified that attention learning in the spatial dimension first and then in the channel dimension finally yielded better experimental results. Park et al. proposed the BAM module, which also considers channel and spatial dimensions, which uses null convolution and uses parallelism to sum two feature maps[9]. In our work, we also use the parallel way to deal with two feature maps, in our opinion, different dimensions they do not exist between the sequential relationship, so we think the parallel way is the best way to deal with two dimensions, and finally confirmed our hypothesis with experimental verification. Compared with the previous Bottleneck Attention Module (BAM), which is optimized in this paper, we add a new shared fully connected layer, and the pooling operation can achieve scale invariance, so we discard the original two 3×3 convolutions and replace them with two pooling (maximum pooling and horizontal pooling) and one 7×7 convolution operation.

However, most of the methods based on deep learning attention mechanism only focus on the channel or space, so that the obtained feature maps can capture some important information and ignore some unimportant information, but we think it is not sufficient for the person re-identification field, and these feature maps cannot cope well with some cases of large pose changes and different camera views. Zhang et al. proposed to learn the attention between each feature node from a global view of the correlation between features and consider the global scope of correlation[10]. Inspired by this, in order to fully establish the relationship between each weight of the feature map, the Self-Relevant Attention (SRA) module is designed in this paper to obtain a more robust feature map by modeling the relationship between each part of the feature map, and then fusing it with the original feature map after coding. In the field of person re-identification, for the obtained feature maps, we believe that smaller feature values should be avoided around higher feature values. Intuitively, if we focus on the face as the most recognizable feature, then the surrounding areas of the face, e.g., the hair, the neck, or the background of the edges, etc., should also be more recognizable. Based on the above considerations, additional feature enhancement operations are added to the final feature map processing part.

Based on the above, we construct a new MLA network that not only focuses more robustly with discriminative features, but also considers the association between feature graph nodes internally. As shown in Fig. 1, our attention model is visualized, in summary, the contribution of our designed MLA network is as follows:

1. We redesign the Bottleneck Attention (BAM) module and added a new fully connected shared module to the BAM module to reduce the parameter operations. So that the improved BAM module focuses more on discriminative features to suppress irrelevant features and generate attention graphs.
2. We design a new Self-relevant Attention (SRA) module, which focuses more on the association between individual feature nodes and models the global structural information, these two modules complement each other and together contribute to pedestrian re-identification.
3. Before the fully connected layer of the network, an additional "average pooling" operation is performed on the feature maps in order to make the resulting feature vectors 'dilatation'.
4. A multi-level attention (MLA) network is introduced in the paper, consisting of a bottleneck attention module (BAM) and a self-relevant attention (SRA) module. These two modules can be freely embedded into any neural network. We conduct extensive experiments on several standard pedestrian re-identification datasets, where the MLA

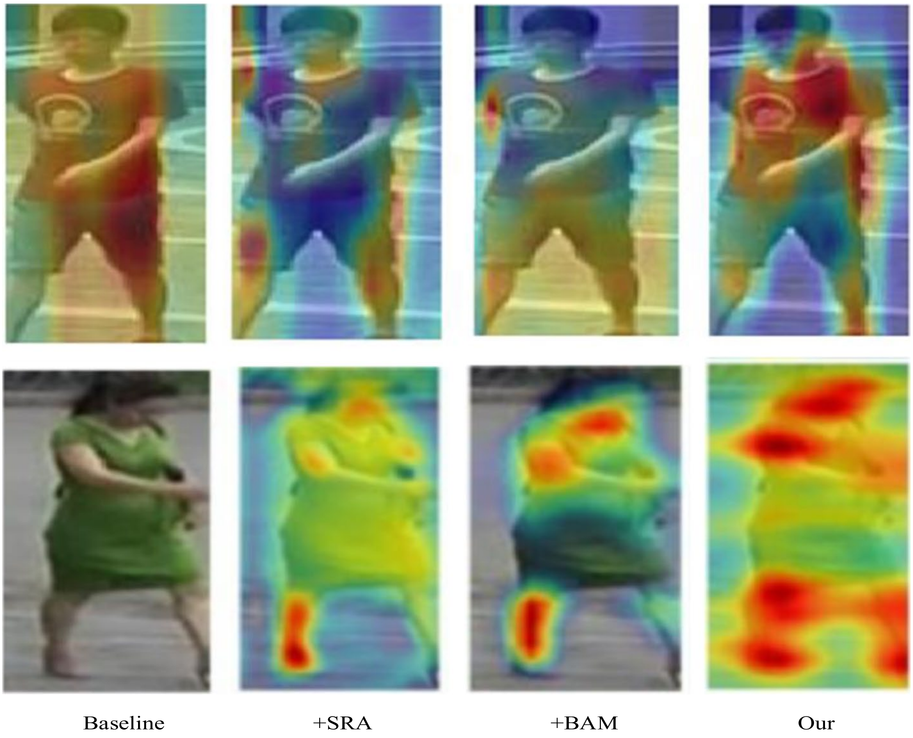


Fig. 1 Attention visualization diagram, darker colors represent higher weights. Baseline represents the baseline network. From the Fig. 1, it can be seen that our proposed network obtains feature maps with more fine-grained focus on the region of interest compared to the baseline network

network significantly outperformed other networks, and validate the model more intuitively through rigorous ablation experiments and visualization.

2 Research works

This section mainly summarizes the research status of person re-identification based on deep learning at home and abroad.

Treating pedestrian re-identification as a specific pedestrian retrieval problem, most of the work uses network architectures designed for image classification, such as the Resnet, Densenet, and Vgg network. With the continuous development of deep learning has greatly boosted the research of computer vision and thus has brought great help to the research of pedestrian re-identification. Deep learning-based pedestrian re-identification has achieved incredible performance over traditional algorithms, and deep learning-based pedestrian re-identification can be divided into representation learning and metric learning[11, 12]. The representation-based learning approach treats pedestrian re-identification as a classification task without considering the similarity between pictures. To improve the performance, some research workers combine the attribute loss of pedestrians with softmax loss to improve the accuracy of experimental results. The metric-based learning approach can be regarded as a clustering task, where different pictures are transformed into features and then the similarity between them is calculated, which can be classified into different loss functions according to different calculation methods[13, 14]. There are also some research workers who train the network with representation learning and metric learning together, which also greatly improves the performance of the network.

2.1 Research on person re-identification based on attention mechanism

Humans can naturally and efficiently find salient regions in complex scenes. Motivated by this situation, attentional mechanisms were introduced into computer vision with the aim of mimicking this aspect of the human visual system. This attention mechanism can be seen as a dynamic weight adjustment process based on the input image features. In recent years, attention mechanisms have played an increasingly important role in computer vision, and Jaderberg et al. proposed the STN network, which introduces a sub-network to predict the affine transformation used to select important regions in the input[15]. Park et al. propose the BAM attention module, which can be integrated with any CNN network[9]. Hu et al. proposed the SENet network, which adaptively predicts potential key features[7]. CBAM[8] proposed by Woo et al. is a lightweight module combining channel attention and spatial attention proposed based SENet, which can be embedded in almost any CNN network to substantially improve model performance while bringing a small amount of computation and number of parameters. In recent years, the emergence of self-attention networks (transformers) has further demonstrated the great potential of attention-based models.

With the appearance of attention models, some researchers have applied attention models to the direction of pedestrian re-identification. Chen et al. proposed HAB and PAB modules, where the HAB module makes it possible to focus on the pedestrian itself while reducing the influence of the background[16]. The PAB module enables the extracted features to be decoupled into multiple groups of different body parts, thus improving robustness to pose changes and occlusions. Huang et al. propose a batch attention (BA) method, which provides a new research method and ideas for pedestrian recognizance based on attention model[17]. The HAG network proposed by Zhang et al. which adopts a hybrid attention module and adds multi-granular operations to fuse high-order and low-level information[18]. Rao et al. propose a counterfactual attentional learning method based on causal reasoning to learn more effective attention maps[19]. Qin et al. proposed Width channel attention Network (WCA) and Depth Channel Attention Network (DCA) [20]. There is also an adaptive weight adjustment strategy to optimize weight coefficients and network parameters. The ABD-Net proposed by Chen et al. consists of a channel attention module (CAM) and a positional attention module (PAM)[21]. Zhang et al. proposed a global attention module (RGA), which considers the relationship with other features for each feature, i.e., considers both global and local information[10]. Si et al. proposed a new Spatial-Driven Network (SDN) to learn special recognition features with rich semantic information from inter-image and intra-image dependencies for person re-identification[22].

Some other researchers have introduced transformer architectures into the field of pedestrian re-identification. For example, Wang et al. proposed the NFormer network[23], which considers the interaction information between images and introduces two other related modules in order to reduce the computational effort. Zhu et al. proposed a dual cross-attentive learning (DCAL) module to coordinate self-attentive learning[24].

2.2 Research on pedestrian re-identification based on pose information

In order to maximize accuracy, some research workers use key points of the human body as auxiliary information to build models. For example, Zheng et al. used CMA to extract the pose points of pedestrians, which were divided into several small image blocks according to the pose points for radiometric transformation, and finally fused the features of the original image and the radiometric image and trained the network using cross-entropy loss[25]. The strategy adopted by Zhao et al. is very similar to that of Zheng et al. except that the Spindle Net proposed by Zhao divides the pose structure of pedestrians into more detailed levels and uses a partial fusion of features before training to train the network[26]. Su et al. proposed that the improved radiation transformation network is called the PTN network, which automatically learns affine transformation parameters[11]. The PABP network proposed by Wang et al. uses the feature maps extracted by the convolutional neural network and the features extracted using the key point extractor for vectorial extrapolation and vectorization, with the intrinsic meaning of activating the appearance features at the corresponding locations[27]. The general idea of these research works with pose information is to use the pose estimation model to derive key points, obtain part regions with semantic information based on the key points, and extract local features for each part region and sometimes combine local features with global features for joint optimization. However, such networks often require training extra models and consume a lot of GPU memory and time, so that these methods are not generally used in industry.

2.3 Other related studies

In pursuit of better performance, some research workers began to explore toward local features. Sun et al. first proposed the idea of horizontal chunking, which simply means that the feature map is divided into six pieces, each of which is trained by the network, and finally a lossy joint optimization model is obtained[26, 28]. Zhang et al. proposed to consider the correlation between different parts after chunking and used graph convolutional networks to assist the training[29]. Luo et al. successively proposed AlignedReID + + network and STNet network to study the work of pedestrian re-identification from the perspective of dynamic chunk alignment and radial transformation to predict the whole-body map, respectively[30, 31]. Some research workers have also studied de-background in the early stage, but as the depth of the network increases, de-background may be detrimental to improve the accuracy of recognition, firstly, because de-background may destroy the pose of the human foreground, and secondly, because it is extremely easy to cause overfitting, so de-contextualization is rare in the current research.

Another research perspective on pedestrian re-identification is based on the Generative Adversarial Network (GAN) approach, where Zhong et al. propose CycleGAN networks to experiment with style shifts between two cameras, and then use label smoothing and finally compute the cross-entropy loss, which is used to eliminate the bias caused by different camera views[32]. The main idea of the network proposed by Wei et al.[33] and Deng et al. [34] are to generate the image of scene A into the style of scene B to improve the generalization ability of the model, the PNGAN network proposed by Qian et al. mainly uses GAN to solve the problem of pose change, where the input is an original image and a pose point image, the generated pose is not the same as the original pose, and the features obtained from different networks are input separately and then fused[35].

3 Methodology

In this section, we first introduce the two network modules and feature map reinforcement method we proposed, followed by the overall network architecture and the loss function. The general framework is shown in Fig. 2. the BAM module generates attention weight maps in both channel and spatial dimensions, respectively, and then sums the generated attention weight maps to obtain the total attention map, and finally this attention map is multiplied with the original features, which can be arbitrarily embedded after each network layer. In this article, we embed the BAM module into layer1 ~ 4. The SRA module is mainly responsible for the feature map to establish its own relationship, first we expand the feature map according to the dimensional direction, then establish the mutual relationship between the feature map (please note that we think the relationship between a and b is not the same as the relationship between b and a), and finally get the self-attentive feature map.

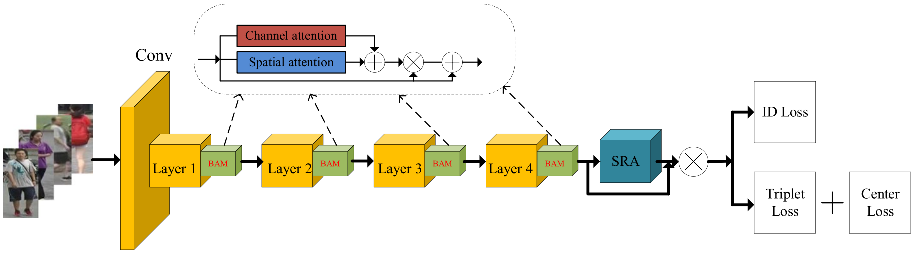


Fig. 2 MLA network architecture. The backbone network uses Resnet50, the BAM modules are placed after the layer of the backbone network, the SRA module is inserted after the last layer as well, and finally the obtained features are used to calculate the triplet loss and center loss, and the features are passed through the FC layer to calculate the ID loss

For the input batch of pedestrian pictures, the first layer is sent to resnet50 (resnet50 network can be divided into 4 layers, each layer contains some convolutional and pooled layers), and the feature map after the first layer is input into the bam module designed in this paper. Each layer of resnet50 contains the BAM module. Finally, the output feature map is input to the SRA module. After the SRA module, the cross entropy loss (ID Loss) and triple loss are calculated respectively. The BAM module is designed to accelerate the neural network's learning of pedestrian features in the picture, and the SRA module is designed to correlate information across the picture (pedestrians should be less correlated with the background, while the head should be more correlated with the upper body and the upper body should be more correlated with the lower body based on the structural characteristics of the human body).

3.1 Feature Map Reinforcement Operation

For the obtained $N \times C \times H \times W$, firstly, the feature map normalization operation will normalize the value to $(0, 1)$, given the threshold a, if the value is greater than the threshold, it is considered that the point as well as the surrounding area is more recognizable features, so for the point around the value less than the threshold with the point for averaging, the example of Fig. 3 is shown. The calculation formula is as follows:

$$X_{i-m,j-n} = (X_{i-m,j-n} + a)/2, (if\, X_{i,j} >= a\,) \tag{1}$$

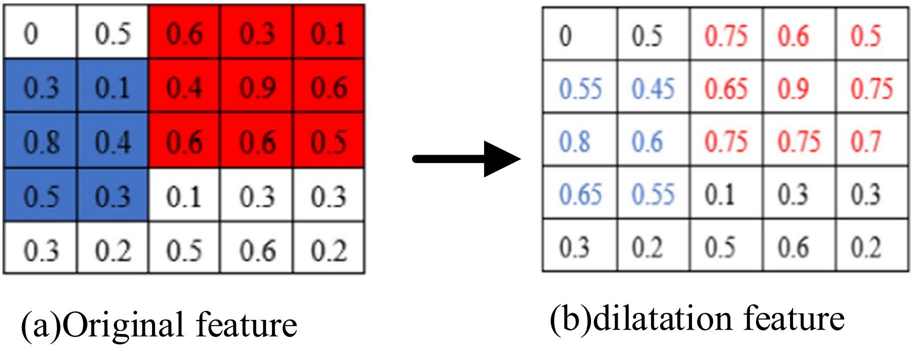
where $i \in (0, H), j \in (0, W), m, n \in \{0, 1\}$, when $m=0,n\neq 0$.

3.2 Bottleneck attention module

The detailed structure of the bottleneck attention module is shown in Fig. 4. For a given input feature tensor the size of the output feature map after the BAM module is:

$$F\prime = F + F \otimes M(F) \tag{2}$$

where $\otimes$ stands for element-by-element multiplication, to design a powerful and efficient module, we compute channel attention $M_C(F) \in \mathbb{R}^C$ and spatial attention $M_S(F) \in \mathbb{R}^{H \times W}$ on two separate branches and then compute the attention characteristics as:



0	0.5	0.6	0.3	0.1
0.3	0.1	0.4	0.9	0.6
0.8	0.4	0.6	0.6	0.5
0.5	0.3	0.1	0.3	0.3
0.3	0.2	0.5	0.6	0.2

(a)Original feature

0	0.5	0.75	0.6	0.5
0.55	0.45	0.65	0.9	0.75
0.8	0.6	0.75	0.75	0.7
0.65	0.55	0.1	0.3	0.3
0.3	0.2	0.5	0.6	0.2

(b)dilatation feature

Fig. 3 Suppose we set $a=0.8$, if there is a value bigger than this threshold then the values around this value are averaged with this value to prevent small values around the big value, so that the feature map obtained is more robust

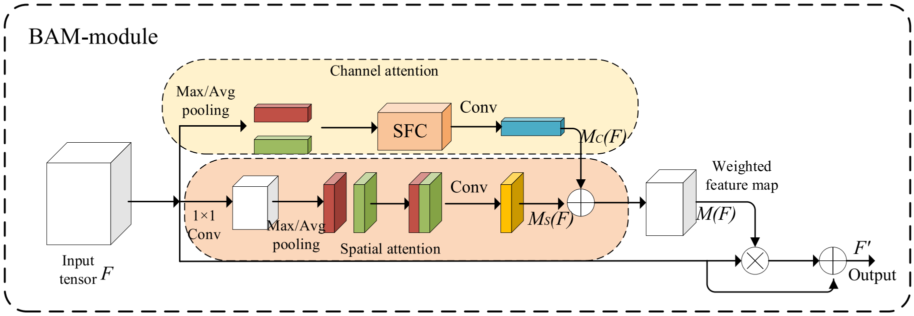


Fig. 4 BAM module architecture. This module contains the channel attention module and the spatial attention module, the spatial attention module and the channel attention module are combined in parallel, and the two are added together to generate a weight map then multiplied with the original feature map to obtain the attention map

$$M(F) = \sigma(M_C(F) + M_S(F)) \tag{3}$$

where σ is the sigmoid function and both branches are resized before summing to $\mathbb{R}^{C\times H\times W}$.

3.2.1 Channel attention module

Shared Fully Connected Module (SFC) The module has two fc layers and its structure is shown in Fig. 5. For the feature map $M_k \in \mathbb{R}^{C\times1\times1}$ input into this module, the size becomes $M_k \in \mathbb{R}^{C/r\times1\times1}$ after the first fully connected layer, and after the second fully connected layer it becomes $M_k \in \mathbb{R}^{C\times1\times1}$ again, where r is the ratio. The main purpose of this module is to reduce the redundant calculation of parameters and prevent degradation of feature maps.

Since each channel contains a specific feature response, the channel attention is focused on "what" is meaningful and "what" is irrelevant. To efficiently compute the channel attention, we compress the spatial dimensionality of the input feature map, and we perform global average pooling and maximum pooling operations on the feature map F to generate two different spatial feature maps:F^C_{avg} and $F^C_{\max}$. Both feature maps are then fed to a fully connected shared network (SFC) to generate our channel attention map $M_c(F) \in \mathbb{R}^C$.The channel attention calculation formula is:

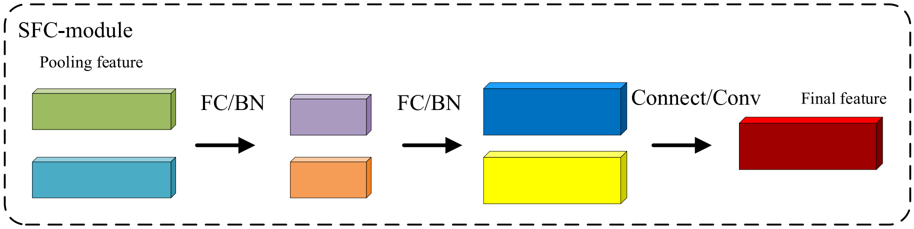


Fig. 5 Shared fully connected module. For the input feature tensor $M_k \in \mathbb{R}^{C\times1\times1}$ after the first fully-connected layer becomes $M_k \in \mathbb{R}^{C/r\times1\times1}$ after the second fully-connected layer becomes $M_k \in \mathbb{R}^{C\times1\times1}$

$$
\begin{aligned}
&M_C(F) = BN(SFC(AvgPool(F)) + SFC(MaxPool(F))) \\
&=BN(W_1(W_0(F^C_{avg} + F^C_{\max}) + b_{_0}) + b_1)
\end{aligned}
\tag{4}
$$

where $W_0 \in \mathbb{R}^{C/r\times C}$,$W_1 \in \mathbb{R}^{C\times C/r}$,$b_0 \in \mathbb{R}^{C/r}$,$b_1 \in \mathbb{R}^{C}$. Note that the weights W_0, W_1 in the SFC are shared.

3.2.2 Spatial attention module

Unlike channel attention, spatial attention uses spatial relationships to generate attention maps, and spatial attention focuses on "where", which is complementary to channel attention. Specifically, we first downscale the features $F \in \mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{C/r\times H\times W}$ using 1×1 convolution, compressing and integrating in the channel dimension. Then apply the average pooling and maximum pooling operations along the channel direction to obtain two feature maps $F^s_{avg} \in \mathbb{R}^{C/r\times H\times W}$,$F^s_{max} \in \mathbb{R}^{C/r\times H\times W}$ and stitch them together to generate a more robust feature, and then perform the convolution operation to produce a two-dimensional attention map, in short, the spatial attention is computed as:

$$
M_S(F) = BN(f^{1\times 1}(f^{7\times 7}(f^{1\times 1}[F^s_{avg};F^s_{\max}])))
\tag{5}
$$

BN stands for batch normalization, $f^{1\times 1}$ stands for convolution kernel with 1×1 convolution and $f^{7\times 7}$ stands for convolution kernel with 7×7 convolution.

After obtaining the attention graphs of the two branches, we combine them to generate the final 3D attention graph $M(F)$, and since the dimensions of the two attention feature graphs do not coincide, we extend both to $\mathbb{R}^{C\times H\times W}$, and then combine them (element-by-element summation or maximal operation), which is experimentally verified in this paper using element-by-element summation, and after summation, we take a sigmoid function to obtain an attention map $M(F)$ in the range of 0 to 1. This 3D attention map is multiplied by the isotopic elements of the input feature map F and then added to the original input feature map to obtain the final feature map $F\prime$.

3.3 Self-relevant attention module

After obtaining the feature attention map of the image, in order to establish the connection between the feature nodes in the feature map, we design the self-relevant attention module, this is shown in Fig. 6. For a given 3-dimensional feature map $\mathbb{R}^{C\times H\times W}$, we expand the feature map so that it becomes a feature map of size $\mathbb{R}^{C\times HW}$, and then use the c-dimensional feature vector at each spatial location as a feature node. We assign a number to the spatial location as $1, 2...N(N = HW)$, and represent the N feature nodes as $x_i \in \mathbb{R}^C$, where $i = 1, 2...N$.

We consider the pairwise relationship (i.e., affinity matrix) from node i to j to be different from that from node j to i. The affinity matrix from node i to node j can be expressed as:

$$
r_{i,j} = f(x_i)^T \xi(x_j),\ r_{i,j} \in \mathbb{R}^{HW\times HW}
\tag{6}
$$

where f and ξ Represents activation functions (including ReLU and BN layers). Similarly, we can obtain the affinity matrix $r_{j,i}$ from node j to node i and then change the dimension of the obtained affinity matrix so that it changes to dimension $\mathbb{R}^{HW\times H\times W}$, and finally let the two stitch together and then reduce the dimension by a 1×1 convolution kernel, please refer to Fig. 6 for more details.

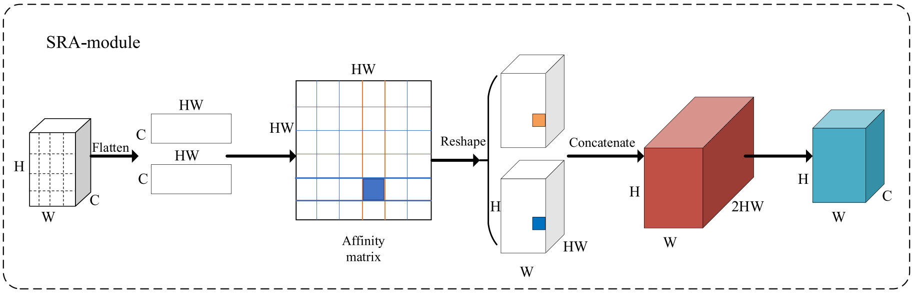


Fig. 6 SRA module architecture. We first flatten the input feature map, and calculate the correlation of a feature point with all locations when calculating the attention of a feature location in order to have the information of the global scope

If the feature tensor of the input self-relevant module is $F\prime$, and the output tensor is A, then the final feature map of the whole backbone network is:

$$F'' = F\prime + F\prime \otimes A \tag{7}$$

where ⊗ represents element-by-element multiplication.

3.4 MLA network architecture and loss function

Our proposed MLA network is shown in Fig. 2, and the backbone network we use is Resnet50. In order to fully consider the best position of these two modules in the backbone network, we conduct sufficient ablation experiments in Chapter 4. Through experimental verification, we embed the BAM module into layer1 ~ 4 and the SRA module into layer4, the optimized BAM module is more able to focus on the region of interest. The SRA module can drive the association of each feature point to further deepen the focus on the region of interest.

The proposed classification loss is used in combination with the cross-entropy loss function, and label smoothing is used to enhance the generalization ability of the network, the loss function is defined as:

$$L_{ID} = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}) \tag{8}$$

where M is the number of categories,y_{ic} is a soft label with a value between (0,1), p_{ic} is the predicted probability that sample i belongs to category c.

The difficult sample mining triplet loss is:

$$L_T = \frac{1}{p \times k}\sum\nolimits_{a \in batch}\left(\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha\right) \tag{9}$$

where p represents the number of pedestrians, k represents each pedestrian has k pictures, $d_{a,p}, d_{a,n}$ represent the feature distance between the sample pictures and the positive and negative samples respectively, taking the positive sample pair with the largest feature distance and the negative sample pair with the smallest feature distance,α represents the bias constant, which is set to 0.3 in this paper. The diagram is shown in Fig. 7.

Center loss[36], which simultaneously learns a center for deep features of each class and penalizes the distances between the deep features and their corresponding class centers, makes up for the drawbacks of the triplet loss. The center loss function is formulated as:

$$L_C = \frac{1}{2}\sum_{i=1}^{m} ||x_i - k_{yi}||_2^2 \tag{10}$$

where yi is the label of the j - th image in a mini-batch. The k_{yi} denotes the yi - th class center of deep features. The m is the number of batch size. The formulation effectively characterizes the intra-class variations. Minimizing center loss increases intra-class compactness. In this article we set the final loss function as:

$$L = L_{ID} + L_T + \beta L_C \tag{11}$$

where β is the weight of the central loss. In our experiments, β is set to 0.0005.

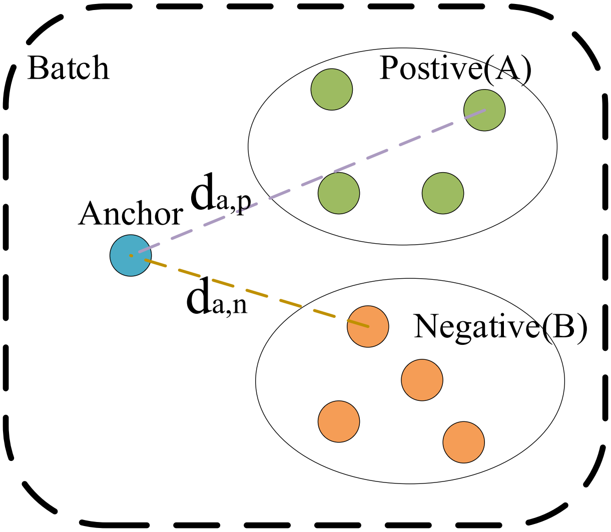
4 Experiment

4.1 Introduction to datasets

Market1501 dataset was collected on the campus of Tsinghua University and images came from six different cameras, one of which was low-pixel. At the same time, the dataset provides training and testing sets. Market1501 contains 32,217 images, the training set contains 12,936 images, and the images are automatically detected and cut by the detector, including some detection errors (close to actual usage). There is a total of 751identities in the training dataset, 750 identities in the test dataset, and an average of 17 training data per class in the training dataset.

CUHK03 dataset was acquired at Chinese University of Hong Kong and images from 2 different cameras. This dataset provides two data sets: machine automatic detection and

Fig. 7 Difficult sample mining triplet loss diagram. Blue circles represent samples to be retrieved, green circles represent positive samples, and orange yellow represents negative samples. $d_{a,p}$, $d_{a,n}$ represent the feature distance between the sample pictures and the positive and negative samples respectively



manual detection. The detection data contains some detection errors, which is closer to the actual situation. The dataset contains a total of 14,097 images of 1,467 pedestrians, The training set has 767 pedestrians, the test set has 700 pedestrians, with an average of 10 training images per person.

MSMT17 dataset (Multi-Scene Multi-Time), covers multiple scenes and multiple time periods. The dataset uses a network of 15 cameras from security on campus, including 12 outdoor cameras and 3 indoor cameras. To capture the raw surveillance video, 4 days with different weather conditions were selected in a month. Three hours of video were captured each day, covering three time periods: morning, noon, and afternoon. Therefore, the total raw video duration was 180 h. The dataset was divided randomly according to a training–testing ratio of 1:3, instead of dividing it equally like the other datasets. This is done to encourage efficient training strategies, due to the expensive nature of labeled data in real applications. The training set contains 1041 pedestrians with a total of 32,621 enclosing frames, while the test set includes 3060 pedestrians with a total of 93,820 enclosing frames. For the test set, 11,659 enclosures were randomly selected as queries, while the other 82,161 enclosures were used as galleries.

4.2 Experiment implementation details

We use the backbone network Resnet50, and the model is pre-trained on ImageNet, the stride of the last convolutional layer is set to 1, the network parameters are jointly optimized using cross-entropy loss and hard sample mining triad loss and central loss, the size of the input image is set to 256×128, each image is normalized, and the data enhancement method is 0.5 random flipping and random cropping of probability (first zoom in on the original image and then crop it). Batchsize is set to 128, and there are 8 different pedestrians in a set of batch size, each pedestrian has 16 different pictures. We use Adam as our optimizer, initial learning rate is set to 0.0003, momentum is set to 0.9, learning rate decay is 0.1, where epoch is 50 and 100 for learning rate decay. The margin of Triplet loss is set to 0.3.

4.3 Ablation experiments

In this paper, we evaluate this method on three standard pedestrian re-identification datasets and compare it with some articles, and the experimental results are shown in Table 1. In the comparison with the simple resnet50 (without any tricks), it can be found that the accuracy improves dramatically after adding on our proposed module in, and in the comparison with the baseline network, it can also be found that the importance of our proposed module. In order to further explore the advantages of the model, we perform a large number of ablation experiments to demonstrate the effectiveness of our proposed attention module.

4.4 Location of BAM attention blocks

In order to verify the effect of the location of the BAM attention module on the model, while keeping the location of the SRA module unchanged, we individually place the location of the BAM module at several locations in Resnet50: layer1, layer2, layer3, layer4, (layer1, layer2), (layer1, layer2, layer3), our (lyer1, layer2, layer3, layer4) and no BAM

Table 1 Results on Market-1501, CUHK03, MMT17 datasets. R = 1 stands for rank-1 results, R = 5 stands for rank-5, mAP stands for average accuracy, – mean no corresponding results are provided in the corresponding article. FR stands for using Feature Map Reinforcement Operation. As it can be seen from the table, the method presented in this paper is superior to most of the most advanced research methods. Since most of these methods do not provide rank-5 results for CUHK03-d and MSMT17 datasets, only rank5 results for Market1501 are compared

Method	Market1501			CUHK03-d		MSMT17	
	R = 1	R = 5	mAP	R = 1	mAP	R = 1	mAP
ALS-net[37]	91.3	96.1	79.2	58.9	52.6	–	–
DU-net[38]	91.6	97.5	75.9	51.6	49.9	–	–
C2F[39]	94.8	97.2	87.7	81.3	84.1	–	–
AlignedReid + + [30]	91.0	–	77.6	60.9	59.7	66.3	40.6
PAT[40]	95.4	–	88.0	–	–	–	–
RGA-SC[10]	96.1	–	88.4	79.6	74.5	80.3	57.5
CNet[41]	95.7	–	88.5	79.4	74.9	–	–
NFormer[23]	95.7	–	93.0	79.0	76.4	80.8	62.2
SRTNet[42]	96.3	–	90.6	92.2	81.6	–	–
Pose-guided[43]	93.5	–	78.6	84.2	93.6	–	–
BSnet[44]	92.5	–	–	87.9	–	71.7	–
baseline[45]	94.5	–	85.9	–	–	–	–
Ours	**96.3**	**98.0**	**87.8**	**78.5**	**73.6**	**79.8**	**55.3**
Ours(+FR)	**96.5**	**98.0**	**87.8**	**81.0**	**78.5**	**81.3**	**61.4**

module as a comparison. We conduct experiments on the Market1501 and CUHK03 datasets, and the results are shown in Fig. 8a, b, the visualization diagram is shown in Fig. 9. From the experimental results, we can clearly see that only when BAM is placed behind all the layers, the final accuracy is higher, and the results do not differ much when placed behind other layers. Because the rank-1 obtained from the Market1501 dataset is relatively high, the rank-1 obtained without the BAM module is even higher than some other ones with the BAM module, but this phenomenon does not occur in the CUHK03 dataset, so we infer that: for the relatively simple dataset, the accuracy obtained from the baseline itself is relatively high, and the BAM module does not play a very obvious role, for those relatively complex datasets, the accuracy obtained from the baseline is relatively low, and the BAM module plays a very obvious role.

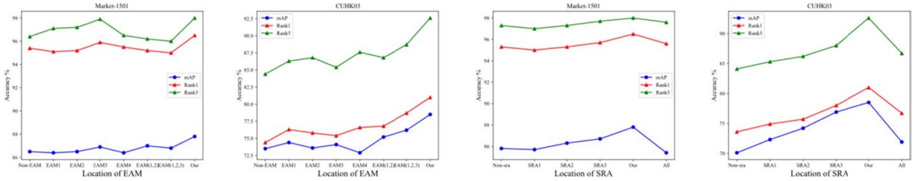


Fig. 8 **a**, **b** represents that effect on experimental results when the BAM module is placed in different positions. From the figure, it can be seen that only when BAM is placed behind all the layers, its accuracy of getting the final is higher. **c**, **d** indicates that effect on experimental results when the SRA module is in different positions. It can be seen from the figure that the highest accuracy is obtained when the SRA module is located at layer4

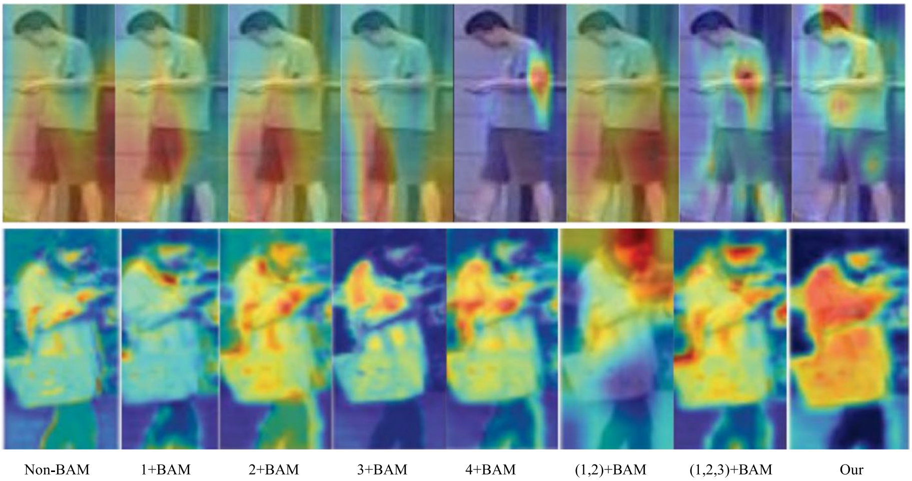


Fig. 9 Attention visualization diagram when the BAM module is in different positions. From the figure, we can see that the BAM module does not play much role when it is placed in the shallow network, and we think that the later convolutional operations are likely to cover the bam module

4.5 Location of the SRA attention module

In order to verify the effect of the position of SRA attention module on the experimental results in the case of keeping the position of BAM module unchanged, we put the position of SRA module separately after layer1, layer2, layer3, layer4 (our) layers, no SRA module and all (layer1, layer2, layer3, layer4), we conduct experiments on Market1501 and CUHK03 dataset, and the experimental results are shown in Fig. 8 c, d, the visualization diagram is shown in Fig. 10. We can see from the experimental results that the impact of the SRA module is still relatively obvious, where the SRA module placed in the last layer gets higher results. We believe that there are two major reasons for this situation: First, because the shallow feature perception field is relatively small, placing the self-relevant attention module in the front end of the network does not learn the more discriminating features well; Second, if the self-relevant attention module is placed in every module, this will cause the degradation of the network and affect the learning of attention.

4.6 The combination of channel and spatial in the BAM module

In the BAM module, the way the attention modules are combined is also an important factor affecting the experimental results. In order to explore the effect of different combinations on the experimental results, we design the following ablation experiments with the combinations shown in Fig. 11, and the experimental results are shown in Table 2. From the experimental results, we can see that the experimental results obtained in parallel form of attention modules are more accurate than those in series, which verifies our conjecture that the priority between attention modules is equal, and the parallel way between attention modules is the most suitable combination. We can also see from the figure that the spatial attention module followed by the channel

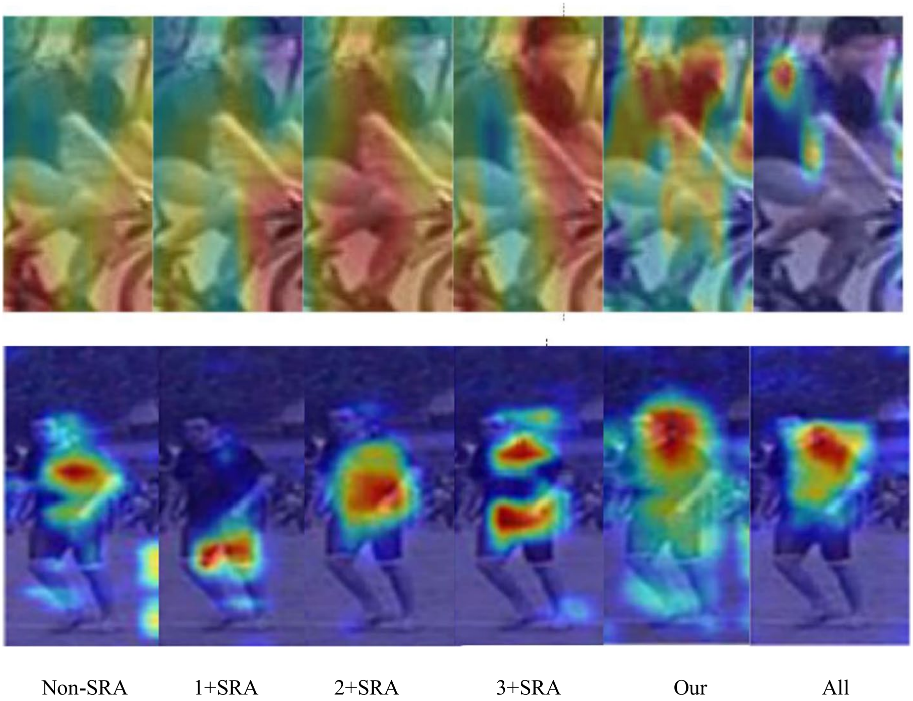


Fig. 10 Visualization of the SRA module at different positions. It can be seen from the figure that the deeper the position of the embedded SRA module is, the more accurate the region of interest of concern is, and if the module is inserted in all layers, it will make the learned obtained region of interest too narrow, instead of being suitable as the final feature map

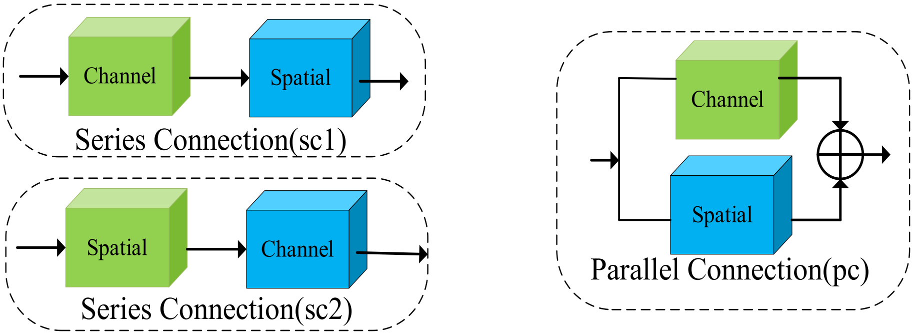


Fig. 11 Combination of channel and space modules in the BAM module in a way that sc1, sc2 are connected in series and pc are connected in parallel

attention module is somewhat more accurate than the channel attention module followed by the spatial attention mod.

Table 2 Effect of different combinations of the channel attention module and spatial attention module in the BAM module on the experimental results. From the table we can see that the parallel way of the two modules finally gives higher experimental results

Dataset	accuracy	pc	sc1	sc2
Market1501	mAP (%)	87.8	87.3	87.0
	Rank-1 (%)	96.5	95.8	94.8
	Rank-5 (%)	98.0	97.4	97.0
CUHK03	mAP (%)	78.5	76.2	75.4
	Rank-1 (%)	81.0	78.6	78.5
	Rank-5 (%)	92.6	90.3	89.8
MSMT17	mAP (%)	61.4	59.8	58.6
	Rank-1 (%)	81.3	79.6	79.8
	Rank-5 (%)	88.7	86.4	86.4

4.7 Ratio in SFC module

The shared fully connected layer in the BAM module is shown in Fig. 5. In order to explore the effect of the scaling ratio in the multilayer perceptron module on the experimental results, we design different ratios r = 4, 6, 8, and 10, and the experimental results are shown in Fig. 12. From the figure, we can clearly see that highest accuracy is achieved when the ratio r = 8. We believe that when the ratio is too small, the network will aggregate features better which is easy to cause overfitting, however, when the ratio is too large, too much information will be lost, which is not conducive to the learning of the network, so the ratio r = 8 is the best ratio state.

4.8 The size of image

To further illustrate the robustness of our proposed model, we consider verifying the robustness of our model by varying the size of the input images. Consider the proportion of human body structure and combine with related paper materials, in this paper we set the size of the pictures as 224 × 224, 128 × 256, 128 × 384, i.e., the size ratios of the pictures are 1:1, 1:2 and 1:3, respectively, and the experimental results are shown in Table 3. From the table we can see that the image size 128 × 256, 128 × 384 finally get the accuracy of

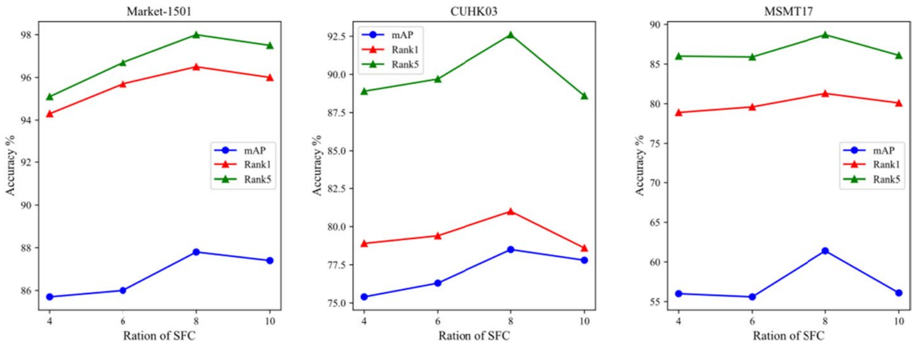


Fig. 12 Effect of the scale size in the SFC module on the experimental results. From the figure, we can clearly see that the highest accuracy is achieved when the ratio r = 8

Table 3 The influence of different input image sizes on experimental results, where the highest accuracy is obtained when the image is inputted W×H = 128×256

W×H		224×224	128×256	128×384
Market1501	Map (%)	87.1	87.8	87.6
	Rank-1(%)	95.8	96.5	96.4
	Rank-5(%)	97.4	98.0	98.1
CUHK03	Map (%)	77.3	78.5	76.9
	Rank-1(%)	79.6	81.0	78.9
	Rank-5(%)	91.5	92.6	92.4
MSMT17	Map (%)	61.6	61.4	59.4
	Rank-1(%)	81.7	81.3	79.6
	Rank-5(%)	87.4	88.7	88.8

the difference is not much (in fact, the image size of 128×384 to get a higher degree of accuracy), are higher than the image size of 224×224, according to our experience of such phenomena are: the larger the input image, the more information contained in the image corresponding to the last obtained feature map, compared to the image ratio of 1: 1, other ratios are more in line with the body shape of pedestrians, which is also a reason to affect the accuracy.

5 Conclusion

In this paper, we propose a new network architecture called multi-level attention model (MLA), which consists of two modules, one is the bottleneck (BAM) attention module and the other is the self-relevant attention (SRA) module. The bottleneck attention module is mainly responsible for generating attention maps, and we further improve it compared to other attention models by adding a shared fully connected layer to our bottleneck attention module, where our module effectively learns what to focus on (or inhibit) and where through two separate pathways. The self-relevant attention(SRA) module, which models the global scope structure information to capture the connections between feature node locations, makes the obtained attention graph more robust by establishing the relationships between individual feature nodes. And based on the ablation experimental results, it is clear that the self-relevant attention module works better after being placed in the deep network layer. Before the fully connecting (FC) layer, we also carried out refined operations on the attention features, in order to make the regional generalization of attention better. In three standard pedestrian re-identification datasets, our experimental results outperform most of the state-of-the-art research results and the module we designed can be arbitrarily embedded into neural networks.

The use of images to estimate the position and direction of the camera can effectively alleviate the problem caused by the inconsistent position of the camera. Theoretically, it can improve the accuracy of pedestrians to re-identify gender. We will continue to work in this area in the future.

Acknowledgements This research is supported by National Natural Science Foundation of China (62101314).

Data Availability Data included in this article can be found at https://github.com/Danyang9999. You can also contact the corresponding author for further information.

References

1. Xiao J, Aggarwal AK, Duc NH et al (2023) A review of remote sensing image spatiotemporal fusion: Challenges, applications and recent trends. Remote Sensing Appl: Soc Environ 32:101005
2. Wu J, Yuan T, Zeng J et al (2023) A Medically Assisted Model for Precise Segmentation of Osteosarcoma Nuclei on Pathological Images, (in eng). IEEE J Biomed Health Inform 27:3982–3993
3. Wu J, Guo Y, Gou F et al (2022) A medical assistant segmentation method for MRI images of osteosarcoma based on DecoupleSegNet. Int J Intell Syst 37:8436–8461
4. Zhou Z, Xie P, Dai Z et al (2024) Self-supervised tumor segmentation and prognosis prediction in osteosarcoma using multiparametric MRI and clinical characteristics. Comput Methods Programs Biomed 244:107974
5. Liu Y, Wang Z, Zhang W et al (2023) DGSN: Learning how to segment pedestrians from other datasets for occluded person re-identification. Image Vis Comput 140:104844
6. Qin W, Huang B, Qin P et al (2022) Learning diverse and deep clues for person reidentification. Image Vis Comput 126:104551
7. Hu J, Shen L, Sun G et al (2018) Squeeze-and-Excitation Networks. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp 7132–7141. https://doi.org/10.1109/cvpr.2018.00745
8. Woo S, Park J, Lee J-Y et al (2018) CBAM: Convolutional Block Attention Module. 15th European Conference on Computer Vision (ECCV) 11211:3–19. https://doi.org/10.1007/978-3-030-01234-2_1
9. Park J, Woo S, Lee JY et al (2018) BAM: Bottleneck Attention Module. British Machine Vision Conference (BMVC), pp. 147–161. http://bmvc2018.org/contents/papers/0092.pdf
10. Zhang Z, Lan C, Zeng W et al (2020) Relation-Aware Global Attention for Person Re-identification. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3183–3192. https://doi.org/10.1109/cvpr42600.2020.00325
11. Su C, Li J, Zhang S et al (2017) Pose-driven Deep Convolutional Model for Person Re-identification. 16th IEEE International Conference on Computer Vision (ICCV), pp. 3980–3989. https://doi.org/10.1109/iccv.2017.427
12. McLaughlin N, del Rincon JM, Miller PC (2017) Person Reidentification Using Deep Convnets With Multitask Learning. IEEE Trans Circuits Syst Video Technol 27:525–539
13. Wei D, Hu X, Wang Z et al (2021) Pose-Guided Multi-Scale Structural Relationship Learning for Video-Based Pedestrian Re-Identification. Ieee Access 9:34845–34858
14. Hou S, Yin K, Liang J et al (2022) Gradient-supervised person re-identification based on dense feature pyramid network. Complex Intell Syst 8:5329–5342
15. Jaderberg M, Simonyan K, Zisserman A et al (2015) Spatial Transformer Networks. 29th Annual Conference on Neural Information Processing Systems (NIPS) 28:2017–2025
16. Chen Y, Wang H, Sun X et al., (2022) Deep attention aware feature learning for person re-Identification, Pattern Recognition,vol. 126. https://doi.org/10.1016/j.patcog.2022.108567
17. Huang Y, Lian S, Hu H (2022) AVPL: Augmented visual perception learning for person Re-identification and beyond, Pattern Recognition, vol. 129. https://doi.org/10.1016/j.patcog.2022.108736
18. Zhang G, Yang J, Zheng Y et al (2021) Hybrid-attention guided network with multiple resolution features for person re-identification. Inf Sci 578:525–538
19. Y. Rao, G. Chen, J. Lu *et al.*, "Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-identification," in *18th IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1005–1014, 2021.
20. Qin W, Huang B, Qin P et al. (2022) Learning diverse and deep clues for person reidentification, Image Vis Comput,vol. 126. https://doi.org/10.1016/j.imavis.2022.104551
21. Chen T, Ding S, Xie J et al (2019) ABD-Net: Attentive but Diverse Person Re-Identification. IEEE/CVF International Conference on Computer Vision (ICCV), pp. 8350–8360. https://doi.org/10.1109/iccv.2019.00844
22. Si T, He F, Wu H et al. (2022) Spatial-driven features based on image dependencies for person re-identification, Pattern Recognition,vol. 124. https://doi.org/10.1016/j.patcog.2021.108462
23. Wang H, Shen J, Liu Y et al (2022) NFormer: Robust Person Re-identification with Neighbor Transformer. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7287–7297. https://doi.org/10.1109/cvpr52688.2022.00715

24. Zhu H, Ke W, Li D et al (2022) Dual Cross-Attention Learning for Fine-Grained Visual Categorization and Object Re-Identification. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4682–4692. https://doi.org/10.1109/cvpr52688.2022.00465
25. Zheng L, Huang Y, Lu H et al. (2019) Pose Invariant Embedding for Deep Person Re-identification, IEEE Trans Image Process, https://doi.org/10.1109/TIP.2019.2910414
26. Zhao H, Tian M, Sun S et al (2017) Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 907–915. https://doi.org/10.1109/cvpr.2017.103
27. Suh Y, Wang J, Tang S et al (2018) Part-Aligned Bilinear Representations for Person Re-identification. Eur Conf Comput Vis (ECCV) 11218:418–437
28. Hu X, Wei D, Wang Z et al., (2021) Hypergraph video pedestrian re-identification based on posture structure relationship and action constraints, Pattern Recognition,vol. 111. https://doi.org/10.1016/j.patcog.2020.107688
29. Zhang Z, Zhang H, Liu S et al (2021) Part-guided graph convolution networks for person re-identification. Pattern Recogn 120:108155–108165. https://doi.org/10.1016/j.patcog.2021.108155
30. Luo H, Jiang W, Zhang X et al (2019) AlignedReID plus plus : Dynamically matching local information for person re-identification. Pattern Recogn 94:53–61
31. Luo H, Jiang W, Fan X et al (2020) STNReID: Deep Convolutional Networks With Pairwise Spatial Transformer Networks for Partial Person Re-Identification. IEEE Trans Multimedia 22:2905–2913
32. Zhong Z, Zheng L, Zheng Z et al (2018) Camera Style Adaptation for Person Re-identification. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5157–5166. https://doi.org/10.1109/cvpr.2018.00541
33. Wei L, Zhang S, Gao W et al (2018) Person Transfer GAN to Bridge Domain Gap for Person Re-Identification. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 79–88. https://doi.org/10.1109/cvpr.2018.00016
34. Deng W, Zheng L, Ye Q et al (2018) Image-Image Domain Adaptation with Preserved Self-Similarity and Domain-Dissimilarity for Person Re-identification. 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 994–1003. https://doi.org/10.1109/cvpr.2018.00110
35. Qian X, Fu Y, Xiang T et al (2018) Pose-Normalized Image Generation for Person Re-identification. 15th Eur Conf Comput Vis (ECCV) 11213:661–678
36. Wen Y, Zhang K, Li Z et al (2016) 2016 A Discriminative Feature Learning Approach for Deep Face Recognition. 14th European Conference on Computer Vision (ECCV) 9911:499–515
37. Jeong D, Park H, Shin J et al., (2020) Uniformity Attentive Learning-Based Siamese Network for Person Re-Identification, Sensors,vol. 20. https://doi.org/10.3390/s20123603
38. Li R, Zhang B, Teng Z et al (2021) A divide-and-unite deep network for person re-identification. Appl Intell 51:1479–1491
39. Zhang A, Gao Y, Niu Y et al (2021) Coarse-to-Fine Person Re-Identification with Auxiliary-Domain Classification and Second-Order Information Bottleneck, in. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2021:598–608
40. Li Y, He J, Zhang T et al (2021) Diverse Part Discovery: Occluded Person Re-identification with Part-Aware Transformer. in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2897–2906. https://doi.org/10.1109/cvpr46437.2021.00292
41. Zhang G, Lin W, Chandran AK et al (2023) Complementary networks for person re-identification. Inform Sci 633:70–84
42. Yang J, Zhang C, Li Z et al. (2023) Discriminative feature mining with relation regularization for person re-identification, Inform Process Manag, vol. 60 https://doi.org/10.1016/j.ipm.2023.103295
43. Khatun A, Denman S, Sridharan S et al., (2023) Pose-driven attention-guided image generation for person re-Identification, Pattern Recognition,vol. 137 https://doi.org/10.1016/j.patcog.2022.109246
44. Chen G, Zou G, Liu Y et al. (2023) Few-shot person re-identification based on Feature Set Augmentation and Metric Fusion, Eng Appl Artif Intell, vol. 125 https://doi.org/10.1016/j.engappai.2023.106761
45. Luo H, Gu Y, Liao X et al (2019) Bag of Tricks and A Strong Baseline for Deep Person Re-identification. in 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1487–1495. https://doi.org/10.1109/cvprw.2019.00190

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.