Mathematical Biosciences and Engineering

MBE, 20(9): 16913–16938.
DOI: 10.3934/mbe.2023754
Received: 09 July 2023
Revised: 09 August 2023
Accepted: 13 August 2023
Published: 25 August 2023

http://www.aimspress.com/journal/MBE

Research article

Pedestrian re-identification based on attention mechanism and Multi-scale feature fusion

Songlin Liu[1,2], Shouming Zhang[1,*], Zijian Diao[1], Zhenbin Fang[2], Zeyu Jiao[2] and Zhenyu Zhong[2]

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China
[2] Institute of Intelligent Manufacturing, GDAS，Guangdong Key Laboratory of Modern Control Technology, Guangzhou 510030, China

* **Correspondence:** Email: zhangsm@stu.kust.edu.com.

Abstract: Existing pedestrian re-identification models generally have low pedestrian retrieval accuracy when encountering factors such as changes in pedestrian posture and occlusion because the network cannot fully express pedestrian feature information. Therefore, this paper proposes a method to address this problem by combining the attention mechanism with multi-scale feature fusion, and combining the proposed cross-attention module with the ResNet50 backbone network. In this way, the ability of the network to extract strong salient features is significantly improved; at the same time, using the multi-scale feature fusion module to extract multi-scale features from different depths of the network, achieving the complementary advantages between features through feature addition, feature concatenation and feature weight selection. In addition, a feature enhancement method and an efficient pedestrian retrieval strategy are proposed to jointly promote the accuracy of pedestrian retrieval from both the training and testing levels. When tested on the occluded pedestrian recognition datasets Partial-REID and Partial-iLIDS, the accuracy of this method reached 70.1% and 65.6% on the Rank-1 indicator respectively, and 82.2% and 80.5% on the Rank-3 indicator respectively. At the same time, it also achieved high recognition accuracy when tested on the Market1501 dataset and DukeMTMC-reid dataset, reaching 95.9% and 89.9% on the Rank-1 indicator respectively, 89.1% and 80.3% on the mAP indicator respectively, and 67% and 46.2% on the mINP indicator respectively. It can be seen that this method has achieved good results in solving the above problems.

Keywords: deep learning; pedestrian re-identification; cross attention; multi-scale features fusion

1. Introduction

Person re-identification (Re-ID) is the process of retrieving a specific person captured by different cameras. Due to its practical importance in surveillance systems, re-identification technology has attracted widespread research and attention, which is of great significance for security monitoring and pedestrian behavior analysis. Traditional pedestrian re-identification methods usually rely on human experience in the design of feature extraction methods, which has certain limitations and deficiencies. However, with the deepening of research on deep learning, network models represented by convolutional neural networks have been widely used in various recognition and detection tasks and have achieved good results. In 2014, a pedestrian re-identification network based on deep learning was first introduced into this field [1], and its main task is to design an effective model to extract pedestrian features with high representational capabilities. According to the different network loss functions, it can be divided into representation learning and metric learning methods.

With the deepening of research on pedestrian re-identification, in addition to representation learning and metric learning methods, there are also methods based on global features, attention mechanisms, video sequences and unsupervised and weakly supervised learning. Wei et al. [2] proposed a global-local alignment feature method, which first roughly divides pedestrians into three parts: head, upper body and lower body, and then learns the original image and local areas separately to obtain more detailed representation features, thereby solving the problem of misaligned pedestrian posture. Sun et al. [3] proposed the PCB method of feature equalization and the RPP strategy which horizontally divides pedestrian features into six blocks and separately learns the local features of different divided areas of pedestrians. Li et al. [4] HA-CNN deep network structure uses soft and hard attention mechanisms to learn global and local features of pedestrians in multiple branches and fuse them together, which can optimize misaligned pedestrian images. Chen Ying et al. [5] proposed a multi-scale learning network combining Convolutional Neural Network (CNN) and TransFormer, effectively utilizing hierarchical features at different scales. Jin Mei et al. [6] AEFC-Net model fully utilizes the differences between hierarchical features and combines asymmetric enhanced attention to make the network pay more attention to the pedestrian area in the image, thereby obtaining sufficient and identifiable pedestrian feature information. In order to solve the problems of spatiotemporal information mining, redundant information suppression and data quality in video pedestrian re-identification, Yang et al. [7] designed a dual-stream dynamic pyramid representation model. Xia et al. [8] used the IMT network to generate target samples, expand the scale of the training dataset and increase its diversity, and then combined the source image with the modality transfer image for training the pedestrian re-identification model to improve cross-modal retrieval performance.

Although the above methods can achieve high recognition accuracy, there are still the following problems that cannot be solved:

- The traditional method of manually extracting features is too complicated and unreasonable, and the amount of computation is relatively large. In addition, the accuracy of pedestrian re-identification models based on traditional methods is not high, making it difficult to achieve the desired results.
- The existing deep learning-based pedestrian re-identification methods also have single feature extraction and insufficient convolutional backbone feature extraction capabilities, and cannot fully and effectively utilize the multi-scale detail information inherent in

pedestrians.

- Due to the insufficient feature expression caused by pedestrian occlusion and small inter-class differences, deep learning-based pedestrian re-identification has not been able to effectively cope with it.

Therefore, in the process of studying deep learning algorithms, this paper uses a residual network combined with an attention mechanism to design a more robust and stronger feature extraction pedestrian re-identification model to achieve full expression of pedestrian feature information, thereby improving recognition efficiency and accuracy. On this basis, further research is conducted on the problem of difficult pedestrian retrieval under special circumstances. The main contributions are as follows:

- Proposed a cross-attention mechanism and embedded it into each convolutional layer of ResNet50 layer by layer to establish long-term dependence between features, allowing the network to pay more attention to some detailed information of pedestrians and make the extracted features more discriminative;
- To make up for the problem that existing pedestrian re-identification methods cannot fully and effectively express pedestrian feature information, a multi-scale feature fusion strategy is proposed. The relationship between features at each layer of the network is fully utilized. Through mutual correlation and feature fusion, an aggregated feature with high representation ability is constructed to obtain stronger fine-grained pedestrian features;
- A random grayscale transformation strategy is proposed. By randomly discarding some color information in the training data, it can effectively solve the problem of difficult matching and mismatching caused by color deviation of pedestrians, and reduce the over-fitting phenomenon of the model;
- For the recognition of noisy pedestrian images with a lot of noise, there is currently a lack of effective processing methods for existing pedestrian re-identification methods. Therefore, a simple and efficient retrieval method is proposed. Side window filtering is used to denoise the noisy images, which can effectively improve the accuracy of retrieval.

The rest of this paper is organized as follows. In Section 2, we briefly introduce the related works of image processing, ResNet and pedestrian re-identification. In Section 3, we describe in detail the proposed pedestrian re-identification method based on attention mechanism and multi-scale feature fusion, including the overall framework, data preprocessing methods, related modules and loss functions. In Section 4, we compare our method with other mainstream methods and show the impressive performance of our method. Finally, in Section 5, we conclude.

2. Related work

In current pedestrian re-identification tasks, most methods directly extract the discriminative features of the entire pedestrian image and use their most prominent discriminative features to identify different pedestrians. This can achieve high accuracy in general scenarios, but with the continuous development of economy and technology, the scenes where pedestrians are located are becoming more and more complex, and the angle of the camera is fixed, resulting in certain occlusion, posture changes

and other phenomena in the captured pedestrian images. In the face of the above phenomena, early pedestrian re-identification solutions proposed mainly used manually designed, better visual features and learned better similarity measures, but both have certain limitations and cannot adapt to large-scale data and complex scene recognition tasks.

With the rapid development of deep learning, many pedestrian re-identification methods have been proposed to solve the above problems. For example, Cheng et al. [9] proposed a general network benchmark for pedestrian re-identification tasks, aiming to extract a discriminative feature that brings pedestrians of the same category closer and those of different categories farther apart; Liu et al. [10] and Wei et al. [11] introduced attention modules into the network model to enhance the model's ability to extract non-salient features; Yan et al. [12] proposed a loss function for pairwise relationships to effectively learn the difference features of pedestrian appearance; Li [13] proposed a multi-task recognition algorithm combining spatial attention and texture features, ingeniously integrating pedestrian attribute features and pedestrian features; other researchers used Generative Adversarial Nets (GAN) [14] to expand the training dataset and enhance the invariance of input changes, but this method often reintroduces new noise, which also limits the performance gain from generated data.

However, the above methods do not consider the attribute that pedestrians themselves have multi-scale information, so the network cannot fully express pedestrian information. Kim et al. [15] proposed a multi-scale attention residual network that extracts features from multiple time scales, improving the feature learning ability of the multi-scale structure, and applied it to fault diagnosis of rotating machines. Wen et al. [16] applied ResNet50 as a feature extractor in fault diagnosis by converting time-domain fault signals into Red-Green-Blue (RGB) image format as the input data type of ResNet50. In the field of medical imaging, Shin et al. [17] applied a residual network to transcranial focused ultrasound simulations using a multivariate merging method. It can be seen that ResNet has been widely used in multiple fields such as image classification, fault diagnosis, and medical imaging. Therefore, using ResNet50 as the backbone network, this paper proposes a pedestrian re-identification method that combines the attention mechanism with multi-scale feature fusion. We evaluated our method in conventional pedestrian re-identification tasks (Market1501, DukeMTMC-reid and MSMT17) and occluded pedestrian re-identification tasks (Partial-REID and Partial-iLIDS). The experimental results show that our method has good performance and even outperforms some mainstream methods.

3. Methods

This paper proposes a pedestrian re-identification network based on attention mechanism and multi-scale feature fusion (AM-MFF), which combines the attention module with the multi-scale feature fusion module. The overall structure of the model is shown in Figure 1, which is mainly composed of a data preprocessing module, main network, cross-attention module and multi-scale feature fusion module.

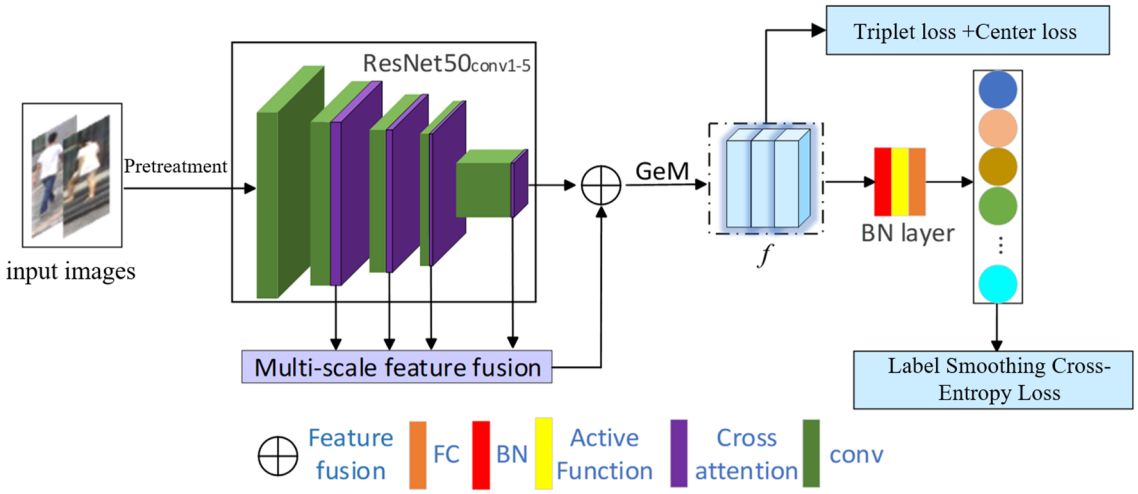


Figure 1. Overall network architecture. Composed of a data preprocessing module, backbone network, cross-attention module, and multi-scale feature fusion module.

First, the pedestrian images are subjected to Random Grayscale Transformation and Side Window Filtering [18] respectively at the input end and retrieval end to enhance the generalization ability of the network and improve the accuracy of network retrieval. Then, the preprocessed pedestrian images are sent into the main network for feature extraction. In this paper, ResNet50 pre-trained on ImageNet is selected as the main network, and the stride of the last down sampling operation is changed from 2 to 1 to increase the spatial size of the output feature map. The fully connected layer of the network is also modified to keep its dimension consistent with the number of identities in the training dataset. To enhance the feature extraction ability of the network, cross-attention modules proposed in this paper are embedded after four convolutional modules conv2_x, conv3_x, conv4_x and conv5_x of ResNet50 respectively. In order to make full use of feature information at different scales in shallow and deep networks, convolution kernels of different sizes are used to obtain feature information at different scales in each level. At the same time, features extracted from each level are progressively added and channel weights are filtered to obtain a total fusion feature f, which can supplement pedestrian detail information. Finally, the GeM pooling strategy is used to further extract fine-grained information on fusion features. Meanwhile, two learnable parameters γ and β are introduced into the BN layer in the network to perform scale transformation and offset on fused pedestrian features and restore data expression ability. After the BN layer, the Sigmoid activation function and fully connected layer are utilized to alleviate the gradient disappearance problem during network training and perform pedestrian classification.

3.1. Data preprocessing

Data preprocessing refers to some processing work before the network is trained. Its main purpose is to optimize the original data to achieve some training objectives, such as improving the quality of original data, obtaining better training results, enhancing the generalization ability of the network and shortening the training time.

3.1.1. Random grayscale conversion

Gao et al. [19] proposed an improved gray-scale transformation algorithm that greatly improved color loss while enhancing color images by processing in RGB space, but the image details were not clear enough. Sun et al. [20] proposed a low-complexity automatic contrast enhancement method that uses high-frequency distribution to estimate the intensity-weighted matrix to control the Gaussian fitting curve and shape the contrast gain distribution. The above methods can adjust the image contrast and brightness and enhance the image by changing the gray-scale values of the image through the gray-scale transformation function. Most of the currently available pedestrian re-identification datasets are composed of photos taken by multiple different cameras in different scenes. A high recognition accuracy can still be obtained through a well-trained neural network. However, in real scenes, the environment is complex and changeable, and the existing datasets are difficult to cover all pedestrian images taken by different cameras, different shooting conditions and different shooting time periods, which lack robustness for other scenes. For example, in some lighting scenes, white and gray, black and dark blue and brown and yellow have certain similarities, and color features are also important discriminant features. Therefore, a data preprocessing method of random grayscale transformation is proposed to weaken the influence of color deviation by doing random grayscale processing on the input image, so as to enhance the generalization ability of the network and prevent overfitting of the model.

In this paper, the probability of random grayscale transformation is set to 0.2. A random rectangular box is generated on the image and a grayscale transformation is performed. The ratio of the input image to the area of the randomly generated rectangular box ranges from 0.02 to 0.4, and the minimum width ratio of the grayscale area is set to 0.3 so that different degrees of grayscale areas can be generated in the pedestrian image. The random grayscale transformation can be expressed by the following equation:

$$I^g = G(I^R) \tag{1}$$

$$rect = RandPosition(I^R) \tag{2}$$

$$I^{kg} = RGT(I^R, I^g, rect) \tag{3}$$

In the equation, I^R is the input pedestrian image, $G(\bullet)$ is the grayscale transformation function, I^g represents the grayscale pedestrian image, $RandPosition(\bullet)$ is the function that generates random rectangular boxes, $RGT(\bullet)$ assigns the pixels in the rectangle corresponding to the I^g image to the input pedestrian image I^R, and I^{kg} represents the pedestrian image after random grayscale transformation.

3.1.2. Side window filtering

The traditional linear filtering processing algorithm usually takes a certain pixel as the center and calculates the adjacent pixels in the window by weighting, that is Eq (4).

$$G_i' = \sum_{j \in \beta_i} w_{ij}\, q_j \tag{4}$$

In Eq (4), β_i is the filtering window with i as the center, w_{ij} represents the weight value, q is the original image pixel value, and G_i represents the output result after the filtering calculation. The weight w will affect the output image after filtering; at the same time, this weighted calculation

method with a certain pixel as the center does not consider the special case of pedestrian edges in the image, which will lead to loss of edge information.

Compared with traditional filtering methods, the use of side window filtering denoising has more advantages. Its core idea is to place the filtered pixel to be processed at a suitable edge position in the filter window, and then generate up, down, left, right, upper left, upper right, lower left and lower right 8 different directions for each pixel to be filtered. For the filtering sub-window, when the edge or corner position of the filtering window is aligned with the pixel point, the best reconstruction result after filtering can be obtained, that is, the final filtering result. This processing method reduces the impact of the filtering window crossing the edge and also retains more edge information. It has good effects in applications such as image smoothing, deblurring and enhancement. The specific calculation steps of the side window filtering algorithm for each pixel are as follows:

- Calculate the filtering result G of 8 different directions of filtering sub-windows, that is, equation 5, where w_{ij} is the weight of the pixel j and n belongs to a set composed of 8 different directions of filtering sub-windows.

$$\begin{aligned} G_n &= \frac{1}{N_n}\sum_{j\in w_i^n} w_{ij}\, q_j \\ N_n &= \sum_{j\in w_i^n} w_{ij} \end{aligned} \tag{5}$$

- Take the best reconstruction result I_m from the 8 filtering results.

According to the above algorithm process, the final output result G'_{SWF} is obtained, where θ, ρ, and r respectively represent the angle between the window and the horizontal line, the position of the target pixel point, and the radius of the filtering window.

$$G'_{SWF} = argmin_{\forall G_i'^{\theta,\rho,r}} \left|\left| q_i - G_i'^{\theta,\rho,r} \right|\right|_2^2 \tag{6}$$

In this paper, side window filtering is used to denoise pedestrian images. Compared with traditional filtering, it will not make the pedestrian image blurred and can better retain the details of the pedestrian. In order to compare the processing effects of traditional filtering and side window filtering, three randomly selected pedestrian images from the query dataset were added with salt and pepper noise respectively, and traditional filtering and side window filtering were performed at the same time. The results are shown in Figure 2. Among them, a is the pedestrian image with salt and pepper noise added, b is the pedestrian image after traditional filtering and c is the pedestrian image processed by side window filtering technology. It can be seen that compared with traditional filtering, the pedestrian image processed by side window filtering technology will not cause too much loss of edge details of pedestrians and image blurring.

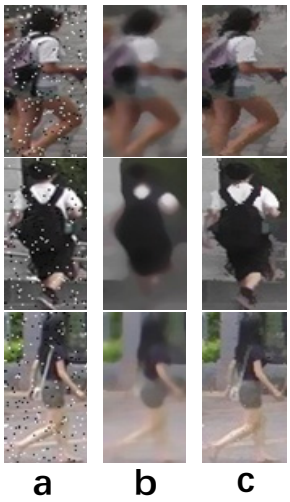


Figure 2. Comparison of traditional filtering and side window filtering effects: a is pedestrian image with added salt and pepper noise; b is pedestrian image processed by traditional filtering; c is pedestrian image processed using the side-window filtering technique.

3.2. Cross-attention mechanism module

Existing pedestrian re-identification methods have proven that adding attention mechanisms to the network can significantly improve the model's recognition accuracy. By using different attention mechanisms, the network's discriminative learning ability can be enhanced by strengthening the connection between different convolutional channels, different body regions, and different images. However, existing attention mechanisms have certain limitations. First, it does not consider that each convolution operation can only process the interaction between local neighborhood information and cannot express data information from farther positions, that is, it lacks effective use of feature position information. Second, it also ignores the special nature of convolution operation, that is, performing the same convolution operation on the same feature map may result in different feature maps. To solve this problem, a cross-attention mechanism (CA) is proposed to associate images with the same identity to enhance the global features of pedestrians and mine more subtle feature information in pedestrian images to reduce differences between images with the same identity.

Currently, self-attention mechanisms have achieved great success in many application fields. As a result, the cross-attention mechanism proposed in this article is an improved method based on the self-attention mechanism. At the same time, it also inherits the ability of the self-attention mechanism to make each pixel perceive all other positions. To better understand the self-attention mechanism, the function definition of self-attention is shown below:

$$SA(X_i) = soft\left(\frac{Q_i \cdot {K_i}^T}{\sqrt{d_K}}\right) \cdot V_i \tag{7}$$

$$Q_i = W_Q X_i\,,\ \ K_i = W_K X_i\,,\ \ V_i = W_V X_i \tag{8}$$

In the equation, X_i represents the input feature map of the i-th convolutional layer, W_Q, W_K, and W_V respectively represent the query, key, and value values obtained by the network from the feature map X_i, d_K represents the dimension of Q_i, K_i, V_i, and soft is a softmax operation.

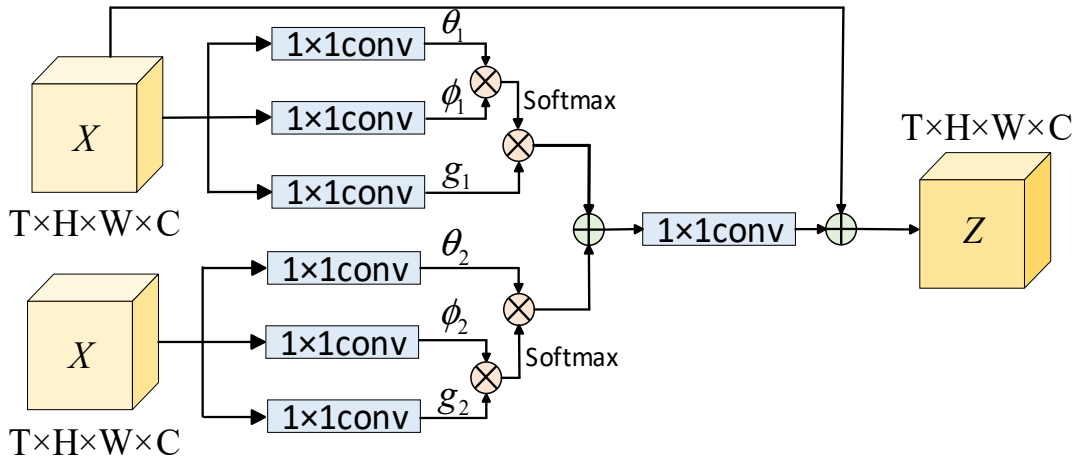


Figure 3. Cross-attention network: T, H, W and C represent the batch size, height, width and number of channels of the feature map, $\otimes$ represents matrix multiplication, $\oplus$ represents element-wise summation.

Figure 3 shows the cross-attention network structure proposed in this article. Specifically, given an input feature map X of size T×H×W×C, after being processed by the cross-attention mechanism, an output feature map Z with the same size as the input feature map X can be obtained. T, H, W and C respectively represent the batch size, height, width, and number of channels of the feature map. $\otimes$ represents matrix multiplication and $\oplus$ represents element summation. The difference from self-attention is that the proposed cross-attention can learn the same image with the same identity and force the network to focus on the different regions in the same image. The specific algorithm process is as follows:

- First, a general cross-attention operation needs to be defined:

$$y_i = soft\left({Q_{i_1}}^T \cdot K_{j_1}\right) \cdot V_{j_1} + soft\left({K_{i_2}}^T \cdot V_{j_2}\right) \cdot Q_{j_2} \tag{9}$$

$$Q_{i_1} = W_{\theta_1} X_i,\ K_{j_1} = W_{\phi_1} X_j,\ V_{j_1} = W_{g_1} X_j \tag{10}$$

$$K_{i_2} = W_{\phi_2} X_i,\ V_{j_2} = W_{g_2} X_j,\ Q_{j_2} = W_{\theta_2} X_j \tag{11}$$

Among them, i_1 and i_2 are the indexes of the output position, which can be the spatial position, time position or spatiotemporal position. The response is obtained by enumerating all positions j. j_1 and j_2 represent all possible positions. W_{θ_1}, W_{ϕ_1}, W_{g_1}, W_{θ_2}, W_{ϕ_2} and W_{g_2} are the weights learned by the network. $Q_{i_1}^T \cdot K_{j_1}$ and $K_{i_2}^T \cdot V_{j_2}$ are used to calculate the similarity between i_1 and j_1, i_2 and j_2 respectively.

- Then, encapsulate the cross-attention operation so that it can be better combined with other network structures:

$$Z_i = CA(X_i) = W_Z y_i + X_i \tag{12}$$

In the equation, W_Z is a learnable weight matrix and $+X_i$ is a residual connection strategy.

3.3. Multi-scale feature fusion module

When performing pedestrian retrieval, the network will use common features of multiple scales to match pedestrians with high similarity. These common features can come from pedestrians themselves, such as clothing, accessories, carry-on items and hairstyles, or from the image acquisition process. For example, the camera's shooting distance is also an important factor affecting the scale and resolution of pedestrian images. When the shooting distance is far away, the scale of the pedestrian image is smaller and the resolution is lower. Some detailed information is not very obvious. When the shooting distance is close, the scale of the pedestrian image is larger and the resolution is higher. Therefore, detailed information will be relatively clear.

In convolutional neural networks, convolutional layers of different depths can obtain feature maps of different scales. This is because deep features have a larger receptive field for feature map acquisition and can capture a wider range of areas. Therefore, they will pay more attention to semantic information, but correspondingly lack spatial geometry and other detailed information; on the contrary, low-level features obtain feature maps with smaller receptive fields, which contain more detailed information, but also more noise. At the same time, a single convolution only uses a fixed-scale convolution kernel to perform convolution operations on the input image, which cannot effectively capture the multi-scale information of the target pedestrian. This also limits the performance of the network.

Given the above problems, that is, the feature expression of pedestrian images at a single scale is not sufficient to support high-precision pedestrian recognition, it is therefore crucial to study a multi-scale feature fusion method. To more comprehensively and effectively use the multi-scale feature information of pedestrians, a new multi-scale feature fusion module is designed. The module includes two components: the Multi-scale Feature Integration (MFI) module and the feature weight selection module. The network structure is shown in Figure 4. The MFI module is used to extract multi-scale feature information from different depths of the network layer, while the feature weight selection module can help us select the most representative and critical features. Specifically: first, multiple-scale feature mutual fusion is performed on the features $X_i(i = 1, 2, 3, 4)$ with input size $H_j \times W_j \times C_j$ $(j = 1, 2, 3, 4)$, which is mainly used to extract multi-scale feature information from different depths of the network, where i and j respectively represent different convolution layers of the backbone network. Then, use the feature weight selection module to select different deep fusion features to obtain a feature map with richer multi-scale feature information. The following section describes in detail the two modules that make up the multi-scale feature fusion.

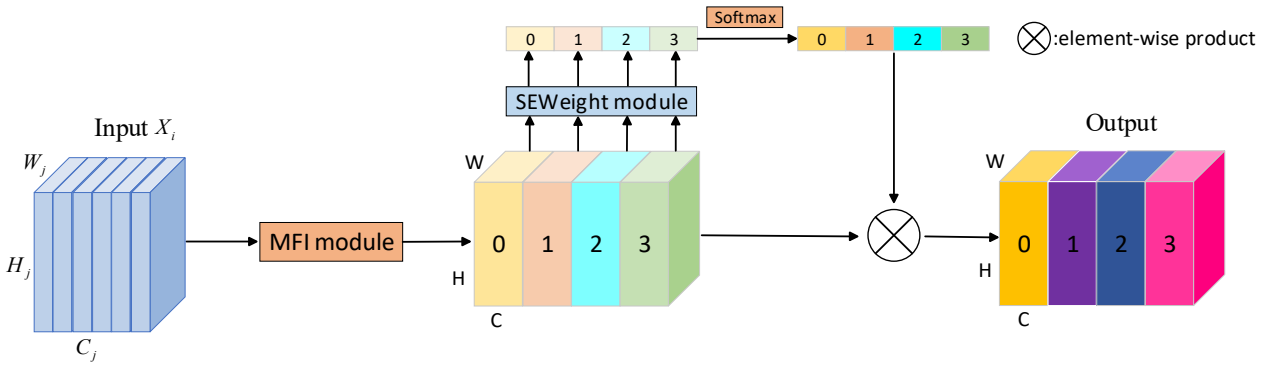


Figure 4. Network structure of the multi-scale feature fusion module. Including the multi-scale feature integration module and the feature weight selection module.

3.3.1. Multi-scale feature interfusion module

MFI can be realized through two steps: multi-scale feature extraction and feature mutual fusion. The following is a detailed introduction.

- Multi-scale feature extraction:

In the multi-scale feature extraction stage, a multi-branch method is used to extract multi-scale information generated by different depth convolution layers, and different sizes of convolution kernels are selected according to the depth of the convolution layer. The size of the convolution kernel is set to 3 × 3, 5 × 5, 7 × 7 and 9 × 9, and the convolution kernel will decrease with the depth of the convolution layer. At the same time, to reduce the computational cost brought by using different convolution kernels, group convolution is introduced in the convolution kernel, and the group size of group convolution is adaptively adjusted according to the change of convolution kernel size. The relationship between them is defined as follows:

$$G = 2^{\frac{K-1}{2}} \tag{13}$$

In the equation, K represents the size of the convolution kernel, and G represents the group size of group convolution. When the convolution kernel K is 3, the group size G of group convolution is 1.

Taking the multi-scale feature extraction of the first layer convolution of the network as an example, the details of feature extraction are shown in Figure 5. Among them, the input feature X_i is the feature obtained by combining the first layer of the ResNet50 residual network and cross-attention. C_1 is the number of channels of input features. Then, parallel processing is used to extract features from input feature X_1 at multiple scales, and four single-type convolution kernel features F_0, F_1, F_2 and F_3 with channel size of 1/2 of the input feature channel are obtained. Finally, feature splicing is performed to obtain fusion feature E_1. The following shows the multi-scale feature extraction process at different depths:

$$F_m = Conv(k_m \times k_m, G_m)(X_i) \quad m = 0,1 \cdots S-1 \quad i = 1,2,3,4 \tag{14}$$

$$E_i = Cat([F_0, F_1, \cdots, F_{S-1}]) \tag{15}$$

Wherein, $k_m = 2 \times (m+1) + 1$ represents the size of the m-th convolution kernel, $G_m = 2^{\frac{k_m - 1}{2}}$ represents the group size of the m-th group convolution, F_m represents the feature map obtained when the convolution kernel size is k_m, E_i represents the multi-scale feature obtained by multi-scale feature extraction of the i-th and i represents the i-th convolution layer of the network, which is inversely proportional to S. For example, when $i = 1$, $S = 4$; when $i = 2$, $S = 3$.

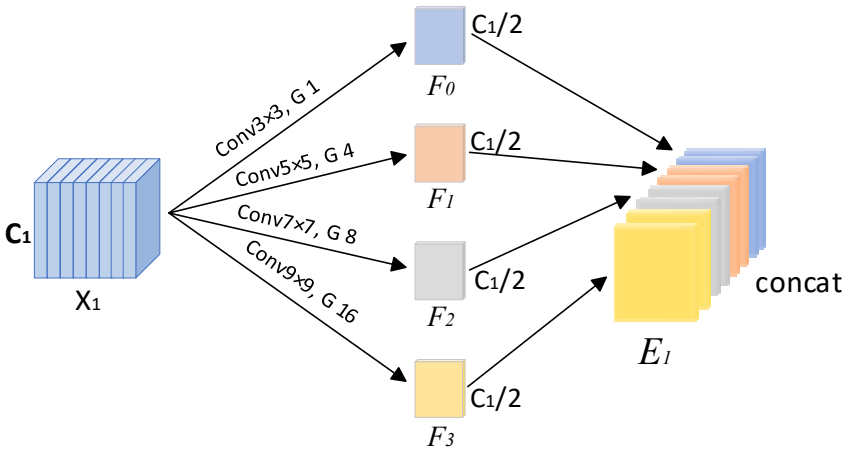


Figure 5. Multi-scale Feature Extraction: X_1 is the feature obtained by combining the first layer of the residual network ResNet50 with the cross-attention, and C_1 is the number of channels of the input feature.

Table 1. Network structure of multi-scale feature extraction.

Layer	Kernel size	Depth	Group
1	3 × 3	128	1
	5 × 5	128	4
	7 × 7	128	8
	9 × 9	128	16
2	3 × 3	128	1
	5 × 5	128	4
	7 × 7	256	8
3	3 × 3	256	1
	5 × 5	256	4
4	3 × 3	512	1

Table 1 shows the convolution kernel scale, depth, and group size of group convolution adopted by different layers of the network. In layer 1, 4 different sizes of convolution kernels are used, and the output channel number of each convolution kernel is set to 128; in layer 2, convolution kernels of sizes 3 × 3, 5 × 5, and 7 × 7 are used, and the output channel numbers are set to 128, 128 and 256 respectively; in layer 3, convolution kernels of sizes 3 × 3 and 5 × 5 are used, and the output channel numbers are both set to 256; in layer 4, only a convolution kernel of size 3 is used, and the output channel number is set to 512. Finally, the multi-scale features obtained at each level are spliced to obtain four multi-scale fusion features E_1, E_2, E_3 and E_4.

- Feature fusion module:

Figure 6 shows the overall process of feature fusion, where Block1, Block2, Block3 and Block4 respectively represent the four new convolutional layers obtained by combining different levels of the residual network ResNet50 with the cross-attention module, and down represents the downsampling operation. The specific description of feature fusion is as follows: the features X_1, X_2, X_3 and X_4 obtained from each new convolutional layer are subjected to multi-scale feature extraction to obtain four multi-scale features E_1, E_2, E_3 and E_4. To better represent the overall features of pedestrians and small-scale detail features, a top-down progressive feature fusion calculation strategy is adopted. At the same time, since the resolution of the fusion feature maps at different levels is not equal, before performing the addition operation, it is necessary to first perform a downsampling operation on the low-level feature map to make the size of the low-level feature map match that of the high-level feature map. Finally, using a progressive fusion calculation strategy, multi-scale features E_1, E_2, E_3 and E_4 are fused to obtain fusion features 0, 1, 2 and 3 containing information of different scales.

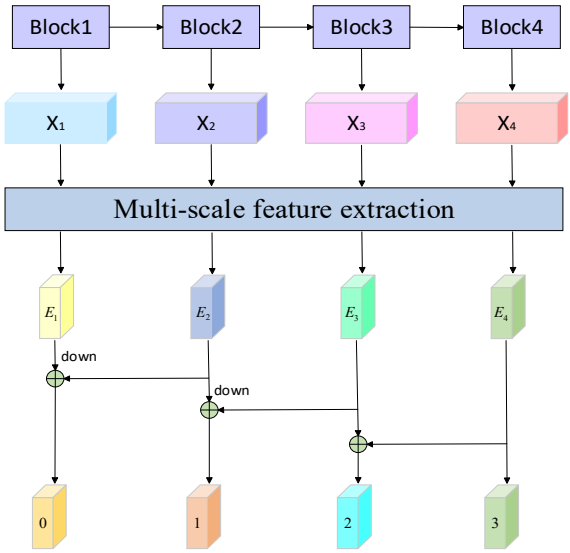


Figure 6. Feature fusion module: Block1, Block2, Block3 and Block4 respectively represent the four new convolutional layers obtained by combining different levels of the residual network ResNet50 with the cross-attention module, and 'down' indicates the downsampling operation.

3.3.2. Feature weight selection module

The feature weight selection module includes two parts: SEWeight weight selection and Softmax weight calibration. The SEWeight module is the most important part of the feature weight selection module, mainly used to obtain the channel weights of different scale features. It essentially belongs to a channel attention mechanism, which allows the network to selectively screen network features, enhance useful features and suppress useless features. The Softmax function is used to recalibrate the channel attention vector to ensure that the weight of each channel is between 0 and 1, and the sum of the weights of all channels is equal to 1.

The implementation process of feature weight selection is specifically introduced as follows: First, use the SEWeight module to extract the channel attention of multi-scale fusion features 0, 1, 2 and 3 respectively to obtain the attention vector on the channel. This allows the network to adaptively focus on the importance of each channel and improve the robustness and representation ability of features. Next, use the Softmax function to recalibrate the attention vector in the channel direction to obtain calibrated weights for multi-scale channels. This allows the model to focus more on channels that make significant contributions to results and weaken channels that are irrelevant or interfere with noise. Finally, perform element-wise multiplication of calibrated weights and corresponding feature maps to achieve weighted fusion of multi-scale feature information. Through this element-wise multiplication operation, a more detailed and comprehensive feature map can be obtained, which contains richer multi-scale feature information.

3.4. Loss function setting

A joint training strategy using label smoothing cross-entropy loss function, triplet loss function, and center loss function is used to calculate the loss of model training. The complete loss function setting is as follows:

$$L_{loss} = L_{tri} + L_{cls} + \beta L_{cen} \tag{16}$$

Wherein, L_{loss} is the total loss of network training, L_{tri}, L_{cls} and L_{cen} are the triplet loss, label smoothing cross-entropy loss and center loss of model training, respectively. β is the weight to balance the center loss, set to 0.0005.

4. Experiment

4.1. Datasets and evaluation indicators

To verify the effectiveness of the model proposed in this paper, experiments were conducted on three standard pedestrian re-identification datasets Market1501, DukeMTMC-reID and MSMT17, and two partial pedestrian re-identification datasets Partial-ReID and Partial-ILIDS.

This paper uses three commonly used evaluation metrics in pedestrian re-identification tasks: cumulative matching characteristic curve, mean average precision and mean negative penalty to evaluate the performance of the network. In addition, different evaluation metrics are selected according to the size and structure of the dataset. On the three standard pedestrian re-identification datasets, Rank-1, mINP and mAP in the cumulative matching characteristic curve are used; on the two partial pedestrian re-identification datasets, only Rank-1 and Rank-3 indicators in the cumulative matching curve are used.

4.2. Comparison with mainstream methods

4.2.1. Comparison of experimental results

In order to verify the performance of the proposed AM-MFF model on Market1501, DukeMTMC-reid, MSMT17, Partial-REID and Partial-iLIDS, it is compared with mainstream

pedestrian re-identification methods such as Bag of Tricks and A Strong Baseline for Deep Person Re-identification (Bag Tricks) [21], Attention Generalized mean pooling with Weighted triplet loss (AGW) [22], Cluster Contrast which stores feature vectors and computes contrastive loss at the cluster level (CCL) [23], Pose-Guided Feature Alignment (PGFA) [24], Augmented discriminative clustering for domain adaptive person re-identification (AD-Cluster) [25] and Fine Grained Spatial Alignment Model (FGSAM) [26]. Through comparative experimental analysis. The experimental results are shown in Table 2.

Table 2. Comparison of experimental results on the Market1501 and DukeMTMC-reid datasets (%).

Methods	Market1501			DukeMTMC-reid		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP
Bag Tricks	94.5	85.9	59.4	86.4	76.4	40.7
AD-Cluster	86.7	68.3	-	72.6	54.1	-
PGFA	91.2	76.8	-	82.6	65.5	-
FGSAM	91.5	85.4	-	85.9	74.1	-
CAD-Net	83.7	-	-	75.6	-	-
UnityStyle	91.8	76.3	-	82.1	65.2	-
SCSN	92.4	88.3	-	91.0	79.0	-
HOA	95.1	85.0	-	89.1	77.2	-
PAT	95.4	88.0	-	88.8	78.2	-
ABD-Net	95.6	88.3	66.2	89.0	78.6	42.1
Fastreid	95.0	87.1	-	88.9	79.0	-
CCL	93.0	82.6	-	85.7	72.8	-
TransReID©	95.0	88.2	-	89.6	**80.6**	-
AGW	95.1	87.8	65.0	89.0	79.6	45.7
Ref	92.4	81.7	-	82.9	69.0	-
PPLR	94.3	84.4	-	-	-	-
AM-MFF (our)	**95.9**	**89.1**	**67.0**	**89.9**	80.3	**46.2**

As shown in the table above, Table 2 shows the comparison results of the method proposed in this paper with various methods such as CAD-Net [27], Unity style transfer for person re-identification (UnityStyle) [28], Salience-guided Cascaded Suppression Network (SCSN) [29], High Order Attention (HOA) [30], Part Aware Transformer (PAT) [31], Fastreid [32], Ref [33] and Part-based Pseudo Label Refinement (PPLR) [34] on the Market1501 and DukeMTMC reid datasets. Our proposed method achieved high recognition accuracy on both datasets, reaching 95.9% and 89.9% on the Rank-1 indicator, 89.1% and 80.3% on the mAP indicator and 67% and 46.2% on the mINP indicator, respectively. Although BagTricks also only used global features for network training, it did not consider that pedestrians may contain more multi-scale feature information. Compared with the proposed method, our method established connections between different deep multi-scale features, resulting in a significant increase in the Rank-1, mAP and mINP indicators of the network, increasing by 1.4%, 3.2% and 7.6%, respectively. On the DukeMTMC-reid dataset, our method achieved the best recognition accuracy on both Rank-1 and mINP indicators, only slightly lower than TransReID© [35] on the mAP indicator. Compared with the ABD-Net [36] network that also uses attention modules, our algorithm improves the ability of the convolutional backbone to extract strong saliency features by

embedding attention modules at each level, increasing by 0.9% on Rank-1, 1.7% on mAP and 4.1% on mINP. According to the above comparison results, it can be seen that our method has certain advantages in performance compared to other mainstream methods on the Market1501 and DukeMTMC-reid datasets.

Table 3. Comparison of experimental results on the MSMT17 dataset (%).

Methods	Rank-1	mAP	mINP
Bag Tricks	63.4	45.1	12.4
SPCL	53.7	26.8	-
CycAS	50.1	16.7	-
MMT	50.1	23.3	-
GCL	45.7	21.3	-
CAP	67.4	36.9	-
ISE	67.6	37.0	-
HDCPD	50.2	24.6	-
CCL	63.3	33.3	-
AGW	68.3	49.3	14.7
AM-MFF (our)	**69.9**	**51.6**	**16.8**

As shown in the table above, Table 3 shows the comparison results of the method proposed in this paper with various mainstream methods such as Self-paced Contrastive Learning (SPCL) [37], Cycle Association (CycAS) [38], Mutual Mean Teaching (MMT) [39], Generative and Contrastive Learning (GCL) [40], Camera Aware Proxies (CAP) [41] and Implicit Sample Extension (ISE) [42] on the MSMT17 dataset, where the Rank-1, mAP and mINP indicators of our method reached 69.9%, 51.6% and 16.8%, respectively, all better than other mainstream methods in the table. Compared with the classic CCL algorithm, our method has increased by 6.6% and 18.3% on Rank-1 and mAP, respectively, with a significant increase in the mAP indicator. Among them, AGW, which uses a non-local attention mechanism, is closest to our method, but it is still 1.6%, 2.3% and 2.1% lower on Rank-1, mAP and mINP indicators, respectively.

In order to further prove that our method is superior and more applicable than other methods, we evaluated and compared the network on two partial pedestrian re-identification datasets. The experimental results are shown in Table 4. Since both Partial-REID and Partial-iLIDS datasets are relatively small in scale and only provide query and gallery libraries, only the evaluation indicators on Rank-1 and Rank-3 are shown. Our method achieved 70.1% and 65.6% on the Rank-1 indicator and 82.2% and 80.5% on the Rank-3 indicator, respectively. Compared with methods such as Deep spatial feature reconstruction (DSR) [43], Bag Tricks, Spatial Transformer Networks for Partial Person Re-identification (STNReID) [44] and AMC+SWM [45], PDVM [46] and Visibility-aware Part-level features Model (VPM) [47] in the table, our method has more significant improvement effects. Among them, the AGW method has the closest recognition accuracy to our method, but it is still lower than our method on both Rank-1 and Rank-3 indicators, thanks to the multi-scale feature fusion module proposed in this paper, which enhances the connection between different scale feature information. From the experimental results in the table, it can be seen that the method proposed in this paper also has advantages on partial pedestrian re-identification datasets, and also shows that our algorithm has good applicability in various practical application scenarios.

Table 4. Comparison of experimental results on the Partial-REID and Partial-iLIDS datasets (%).

Methods	Partial-REID		Partial-iLIDS	
	Rank-1	Rank-3	Rank-1	Rank-3
DSR	50.7	70.0	58.8	67.2
Bag Tricks	62.0	74.0	58.8	73.9
STNReID	61.3	76.8	43.7	62.6
AMC+SWM	37.3	46.0	21.0	32.8
PDVM	43.3	-	-	-
VPM	67.7	81.9	65.5	74.8
AGW	69.7	80.0	64.7	79.8
AM-MFF (our)	**70.1**	**82.2**	**65.6**	**80.5**

4.2.2. Visualization of experimental results

From the experimental results in Tables 4–6, it can be seen that all performance indicators of our algorithm have achieved high recognition accuracy. In order to more clearly and intuitively show the recognition effect of our algorithm, multiple pedestrians to be retrieved are selected from the query library of the Market1501 dataset for visualization, and the top ten retrieval results are shown. The results are shown in Figure 7. The retrieval images with red borders in the figure indicate incorrect query results, and the unmarked retrieval images are correct retrieval results. From the query results, it can be seen that our algorithm still has good recognition effects on partial pedestrian images, overall pedestrian images, blurred pedestrian images, occluded pedestrian images and pedestrian images with complex spatial information. Only the 10th sequence in the second group was wrong, which once again proves the effectiveness of the pedestrian re-identification method proposed in this paper.

Figure 7. Visualization of recognition effect. Red borders indicate incorrect retrieval query results, while unmarked results are correct retrieval results.

4.3. Experimental results and analysis

4.3.1. Ablation experiment

To verify the effectiveness of each component module in the network, ablation experiments were performed on the components separately on the Market1501 and DukeMTMC-reID datasets. The Rank-1, mAP and mINP indicators of the experimental results are shown in Table 5. Among them, Baseline is the result obtained without adding any components, CA represents the cross-attention mechanism, MFF represents the multi-scale feature fusion module and RG represents random grayscale conversion.

Table 5. Ablation experiment (%).

Methods	Market1501			DukeMTMC-reID		
	Rank-1	mAP	mINP	Rank-1	mAP	mINP
Baseline	94.2	86.4	60.1	86.4	76.4	41.2
Baseline+CA	95.0	87.0	61.9	88.2	78.0	42.3
Baseline+MFF	94.9	87.5	63.8	88.0	78.0	43.4
Baseline+RG	94.6	86.7	61.1	86.9	76.9	41.8
Baseline+CA+MFF	95.6	88.7	66.6	89.5	79.6	45.5
Baseline+CA+MFF+RG	**95.9**	**89.1**	**67.0**	**89.9**	**80.3**	**46.2**

From the results in Table 5, it can be seen that on the basis of Baseline, the Rank-1, mAP, and mINP indicators have all improved after adding the cross-attention mechanism (CA), multi-scale feature fusion module (MFF) and random grayscale conversion (RG) proposed in this paper. Compared with using only one of the components, using all three components proposed can achieve the best results for the network. On the Market1501 dataset, the Rank-1, mAP and mINP indicators of the AM-MFF network reached 95.9%, 89.1% and 67%, respectively, an increase of 1.7%, 2.7% and 6.9% compared to baseline. On the DukeMTMC-reID dataset, the Rank-1, mAP and mINP indicators were 89.9%, 80.3% and 46.2, respectively, an increase of 3.5%, 3.9% and 5% compared to baseline.

Further analysis of the experimental results shows that when adding a cross-attention mechanism, on the DukeMTMC-reID dataset, the Rank-1, mAP and mINP indicators increased by 1.8%, 1.6% and 1.1%, respectively, while on the Market1501 dataset there was only a small increase. Among them, on the Market1501 and DukeMTMC-reID datasets, when adding a multi-scale feature fusion module again, the improvement effect of the mINP indicator is significant, increasing by 4.7% and 3.2% respectively, fully demonstrating the effectiveness of multi-scale feature fusion. In summary, the experimental results show that the proposed AM-MFF network can effectively improve pedestrian re-identification performance.

4.3.2. Verification of the effectiveness of random grayscale conversion

To further demonstrate the superiority of our method in instance retrieval, we compared the instance retrieval results of RGB and grayscale images under different training conditions on the Market1501 dataset. The results are shown in Figure 8, where red rectangles indicate incorrect retrieval results and unmarked results are correct retrieval results. The first and third rows are RGB images, while the second and fourth rows are grayscale images. From the retrieval results, it can be seen that

adding random grayscale conversion to the network can alleviate the color deviation problem between the query set and the gallery set. For some pedestrians to be retrieved, if their inherent color information is not considered, their retrieval results will be relatively better, further confirming that random grayscale conversion is effective.

Figure 8. Comparison of retrieval results between grayscale images and RGB images: The first and third rows are RGB images, the second and fourth rows are grayscale images.

4.3.3. Verification of the effectiveness of side window filtering for denoising

Considering that the proposed efficient retrieval method is mainly used to process pedestrian images with more noise in the pedestrian re-identification system, in order to more comprehensively evaluate the effectiveness of the efficient retrieval method, a differentiated verification method is used to test the effectiveness of the method. Specifically, two different levels of salt and pepper noise are added to the pedestrian images to be retrieved, resulting in images a and d. Then, traditional filtering and side-window filtering are used respectively to denoise pedestrian images a and d with added noise, resulting in denoised images b, e and c, f. Then, images a, b, c, d, e, f are retrieved respectively to obtain the top ten retrieval results, as shown in Figure 9. The red box indicates incorrect retrieval results and no frame indicates correct retrieval results.

From the retrieval results in Figure 9, it can be seen that for the pedestrian image a with less noise, after traditional filtering and side-window filtering, the retrieval accuracy is better than that of the retrieval result without denoising. For the pedestrian image after side-window filtering denoising, it can more accurately retrieve the correct pedestrian image. On the contrary, for the pedestrian image d with more noise, traditional filtering fails and cannot retrieve the correct pedestrian image. After side-window filtering denoising, it can still better retrieve the correct pedestrian image, proving that side-window filtering can effectively improve the accuracy of pedestrian retrieval.

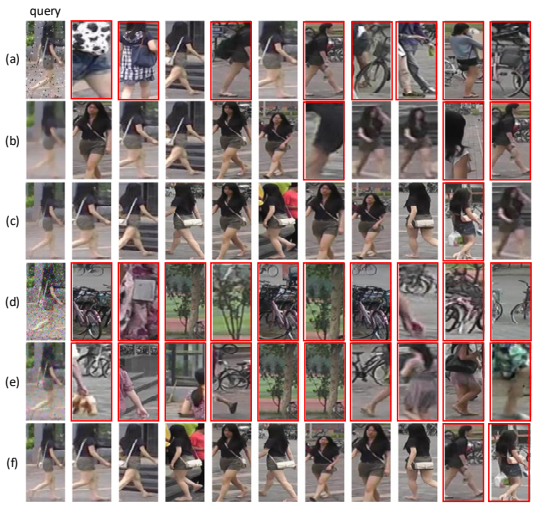
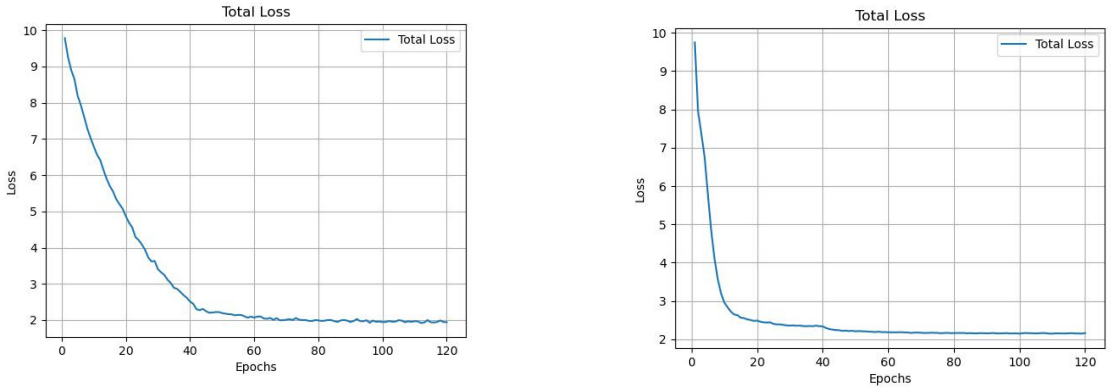


Figure 9. Example comparison and retrieval results. Red boxes indicate incorrect retrieval results, while unboxed results are correct retrieval results.

4.3.4. Comparison of loss curves

In order to further analyze the performance of the method proposed in this paper and baseline, the differences between the two methods are indirectly analyzed by visualizing the loss curve, and the results are shown in Figure 10. The vertical axis in the figure represents the total loss value generated by network training in each training cycle, that is, the sum of the label smoothing cross-entropy loss value, triplet loss value, and center loss value of the network in this cycle. The horizontal axis represents the training cycle (Epoch) of the network. From the comparison chart, it can be seen that the training loss function curve of the method proposed in this paper is smoother and the network convergence speed is faster than that of baseline, which means that the training effect of the proposed model is relatively better and the experimental parameter settings are relatively reasonable. It also indirectly indicates that the proposed method has certain advantages.



(**a**) (**b**)

Figure 10. Comparison of loss curves between baseline and the network in this paper: (a) Baseline's loss curve; (b) The loss curve of this network.

4.3.5. Experimental analysis of random grayscale conversion probability

In order to explore the impact of different random grayscale conversion probabilities on network performance, multiple experiments were conducted on the Market1501 dataset, and the average values of the results of each group of experiments were taken. The experimental results are shown in Table 6. From the results in the table, it can be seen that when the random grayscale conversion probability is set to 0.2, the network achieves the best result, with mINP, mAP and Rank-1 reaching 67%, 89.1% and 95.9%, respectively. Compared with the case without random grayscale conversion, the method proposed in this paper has improved by 0.4%, 0.4% and 0.3% on mINP, mAP and Rank-1 indicators, respectively. As the conversion probability increases, mINP, mAP and Rank-1 indicators all fluctuate to varying degrees. Only when the conversion probability is 0.2 or 1 do all three indicators exceed the performance indicators without random grayscale conversion.

Table 6. Experimental results of the network under different random conversion probabilities (%).

Transition probability	mINP	mAP	Rank-1
0	66.6	88.7	95.6
0.2	**67.0**	**89.1**	**95.9**
0.4	66.5	88.6	95.6
0.6	66.2	89.0	95.7
0.8	66.6	88.9	95.2
1	66.9	**89.1**	95.8

According to Table 3, we can see that when the random conversion probability increases from 0 to 0.2, mINP, mAP and Rank-1 all show an upward trend, and the performance is best when the conversion probability is 0.2. When the conversion probability increases from 0.2 to 1, mINP, mAP and Rank-1 usually show a trend of first decreasing and then increasing, and the performance is best when the conversion probability is 1. Based on the experimental results in Table 3, it is ultimately determined that setting the probability of random grayscale conversion to 0.2 can train the network with the best performance.

5. Conclusions

In this paper, we propose the AM-MFF network, which uses the proposed cross-attention mechanism to enhance the network's feature extraction ability and combines the multi-scale feature fusion module to establish the connection between the low-level feature information and high-level feature information of the network, enabling the model to effectively distinguish image details. At the same time, we perform random grayscale transformation on the training dataset to improve the network's ability to cope with color deviation in images and enhance the network's generalization ability. Moreover, for pedestrian re-identification systems, we propose an efficient retrieval method that performs side-window filtering on noisy pedestrians to be retrieved, which can improve retrieval accuracy to a certain extent. Experiments show that our algorithm achieves 70.1% and 65.6% on the Rank-1 indicator when validated on the occlusion pedestrian recognition datasets Partial-REID and Partial-iLIDS, respectively, and 82.2% and 80.5% on the Rank-3 indicator, respectively. When tested on the Market1501 dataset, DukeMTMC-reid dataset, and MSMT17 dataset, it also achieved high recognition accuracy, reaching 95.9%, 89.9% and 69.9% on the Rank-1 indicator, respectively, 89.1%,

80.3% and 51.6% on the mAP indicator, respectively, and 67%, 46.2% and 16.8% on the mINP indicator, respectively. Thus, our algorithm shows good performance on three conventional pedestrian re-identification datasets and two partial datasets, achieving good results in solving pedestrian posture changes and occlusion problems, with broader applicability and superiority.

The experimental results show that the method proposed in this paper has achieved good results in the task of pedestrian re-identification, and has solved some of the problems existing in current research to a certain extent. However, there are still some limitations and shortcomings in practical applications: 1). Not all pictures have noise, and there may also be situations such as blurring and low resolution. Our method cannot adaptively correct these images and perform corresponding adaptive processing to improve retrieval accuracy; 2). Our pedestrian re-identification is carried out on several commonly used public datasets, which can improve the accuracy of the algorithm to a certain extent. However, these datasets are small in scale and cannot cover pedestrian images in real scenarios, which has certain limitations and cannot better meet the needs of practical applications; 3). The pedestrian re-identification method in this paper does not take into account the issue of pedestrian privacy and security. In order to protect user privacy and data security, deep learning technology can be used to encrypt pedestrian images. The subsequent pedestrian re-identification task can improve the accuracy of the algorithm by solving the above problems and meet the needs of practical applications.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This research was funded by Guangzhou Science and Technology Plan Project (202007040007), Guangdong Basic and Applied Basic Research Fund Project (2022A1515110007), Guangdong Basic and Applied Basic Research Fund Project (2023A1515012869).

Conflict of interest

The authors declare there is no conflict of interest.

References

1. D. Yi, Z. Lei, S. C. Liao, S. Z. Li, Deep metric learning for person re-identification, in *International Conference on Pattern Recognition (ICPR)*, (2014), 34–39. https://doi.org/10.1109/ICPR.2014.16
2. L. Wei, S. Zhang, H. Yao, W. Gao, Q. Tian, GLAD: Global–local-alignment descriptor for scalable person re-identification, *IEEE Trans. Multimedia*, **21** (2018), 986–999. https://doi.org/10.1109/TMM.2018.2870522
3. Y. Sun, L. Zheng, Y. Yang, Q. Tian, S. Wang, Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline), in *Lecture Notes in Computer Science*, Springer, 2018. https://doi.org/10.1007/978-3-030-01225-0_30

4. W. Li, X. T. Zhu, S. G. Gong, Harmonious attention network for person re-identification, *arXiv preprint*, (2018), arXiv:1802.08122. https://doi.org/10.48550/arXiv.1802.08122
5. C. Ying, K. Cheng, Pedestrian re-identification method based on multi-scale learning of CNN and TransForme (in Chinese), *J. Electron. Inf. Technol.*, **45** (2023), 2256–2263. https://doi.org/10.11999/JEIT220601
6. M. Jin, Y. Y. Li, X. J. Hao, M. Yang, L. G. Zhang, Pedestrian re-identification method based on asymmetric enhanced attention and feature cross fusion (in Chinese), *Acta Metrol. Sin.*, **43** (2022), 1573–1580. https://doi.org/10.3969/j.issn.1000-1158.2022.12.08
7. X. Yang, L. C. Liu, N. N. Wang, X. Gao, A two-stream dynamic pyramid representation model for video-based person re-identification, *IEEE Trans. Image Process.*, **30** (2021), 6266–6276. https://doi.org/10.1109/TIP.2021.3093759
8. D. X. Xia, H. J. Liu, L. L. Xu, L. Wang, Visible-infrared person re-identification with data augmentation via cycle-consistent adversarial network, *Neurocomputing*, **443** (2021), 35–46. https://doi.org/10.1016/j.neucom.2021.02.088
9. D. Cheng, Y. H. Gong, S. P. Zhou, J. Wang, N. Zheng, Person re-identification by multi-channel parts-based cnn with improved triplet loss function, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (2016), 1335–1344. https://doi.org/10.1109/CVPR.2016.149
10. Z. Y. Liu, P. P. Wan, Feature extraction method for pedestrian re-identification based on attention mechanism (in Chinese), *Comput. Appl.*, **40** (2020), 672–676. https://doi.org/10.11772/j.issn.1001-9081.2019081356
11. Z. W. Wei, D. Qu, C. Liu, Feature extraction method for pedestrian re-identification based on connection attentio (in Chinese), *Comput. Eng.*, **48** (2022), 220–226. https://doi.org/10.19678/j.issn.1000-3428.0061884
12. C. Yan, G. S. Pang, X. Bai, C. Liu, X. Ning, L. Gu, et al., Beyond triplet loss: person re-identification with fine-grained difference-aware pairwise loss, *IEEE Trans. Multimedia*, **24** (2021), 1665–1677. https://doi.org/10.1109/TMM.2021.3069562
13. J. Li, Pedestrian re-identification enhanced by combining attention and texture features (in Chinese), *Comput. Sci. Explor.*, **16** (2022), 661–668. https://doi.org/10.3778/j.issn.1673-9418.2010046
14. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, *Commun. ACM*, **63** (2020), 139–144. https://doi.org/10.1145/3422622
15. H. Kim, C. Park, C. Suh, M. Chae, H. Yoon, B. Youn, MPARN: multi-scale path attention residual network for fault diagnosis of rotating machines, *J. Comput. Des. Eng.*, **10** (2023), 860–872. https://doi.org/10.1093/jcde/qwad031
16. L. Wen, X. Y. Li, L. Gao, A transfer convolutional neural network for fault diagnosis based on ResNet50, *Neural Comput. Appl.*, **32** (2020), 6111–6124. https://doi.org/10.1007/s00521-019-04097-w
17. M. Shin, Z. Peng, H. Kim, S. Yoo, K. Yoon, Multivariableincorporating super-resolution residual network for transcranial focused ultrasound simulation, *Comput. Methods Programs Biomed.*, **237** (2023), 107591. https://doi.org/10.1016/j.cmpb.2023.107591
18. H. Yin, Y. H. Gong, G. Qiu, Side window filterin, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 8758–8766. https://doi.org/10.1109/CVPR.2019.00896

19. H. Gao, W. Zeng, J. Chen, An improved gray-scale transformation method for pseudo-color image enhancement, *Comput. Opt.*, **43** (2019), 78–82. https://doi.org/10.18287/2412-6179-2019-43-1-78-82
20. X. W. Sun, Q. S. Xu, L. Zhu, An effective Gaussian fitting approach for image contrast enhancemen, *IEEE Access*, **7** (2019), 31946–31958. https://doi.org/10.1109/ACCESS.2019.2900717
21. H. Luo, Y. Z. Gu, X. Y. Liao, S. Lai, W. Jiang, Bag of tricks and a strong baseline for deep person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (2019), 1487–1495. https://doi.org/10.1109/CVPRW.2019.00190
22. M. Ye, J. B. Shen, G. J. Lin, T. Xiang, L. Shao, S. Hoi, Deep learning for person re-identification: A survey and outlook, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2021), 2872–2893. https://doi.org/10.1109/TPAMI.2021.3054775
23. Z. Z. Dai, G. Y. Wang, W. H. Yuan, X. Liu, S. Zhu, P. Tan, Cluster contrast for unsupervised person re-identification, *arXiv preprint*, (2022), arXiv:2103.11568. https://doi.org/10.48550/arXiv.2103.11568
24. J. Miao, Y. Wu, P. Liu, Y. Ding, Y. Yang, Pose-guided feature alignment for occluded person re-identification, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 542–551. https://doi.org/10.1109/ICCV.2019.00063
25. Y. P. Zhai, S. J. Lu, Q. X. Ye, X. Shan, J. Chen, R. Ji, et al., Ad-cluster: Augmented discriminative clustering for domain adaptive person re-identification, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 9021–9030. https://doi.org/10.1109/CVPR42600.2020.00904
26. Q. Q. Zhou, B. N. Zhong, X. Y. Lan, G. Sun, Y. Zhang, B. Zhang, et al., Fine-grained spatial alignment model for person re-identification with focal triplet loss, *IEEE Trans. Image Process.*, **29** (2020), 7578–7589. https://doi.org/10.1109/TIP.2020.3004267
27. Y. J. Li, Y. C. Chen, Y. Y. Lin, X. Du, Y. Wang, Recover and identify: A generative dual model for cross-resolution person re-identification, in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 8090–8099. https://doi.org/10.1109/ICCV.2019.00818
28. C. Liu, X. J. Chang, Y. D. Shen, Unity style transfer for person re-identification, in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 6887–6896. https://doi.org/10.1109/CVPR42600.2020.00692
29. X. S. Chen, C. M. Fu, Y. Zhao, F. Zheng, J. Song, R. Ji, et al., Salience-guided cascaded suppression network for person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2020), 3297–3307. https://doi.org/10.1109/CVPR42600.2020.00336
30. B. H. Chen, W. H. Deng, J. N. Hu, Mixed high-order attention network for person re-identification, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 371–381. https://doi.org/10.1109/ICCV.2019.00046
31. Y. L. Li, J. F. He, T. Z. Zhang, X. Liu, Y. Zhang, F. Wu, Diverse part discovery: Occluded person re-identification with part-aware transformer, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 2897–2906. https://doi.org/10.1109/CVPR46437.2021.00292

32. L. X. He, X. Y. Liao, W. Liu, X. Liu, P. Cheng, T. Mei, Fastreid: A pytorch toolbox for general instance re-identification, *arXiv preprint*, (2020), arXiv:2006.02631. https://doi.org/10.48550/arXiv.2006.02631
33. D. Cheng, J. Y. Zhou, N. N. Wang, X. Gao, Hybrid dynamic contrast and probability distillation for unsupervised person Re-Id, *IEEE Trans. Image Process.*, **31** (2022), 3334–3346. https://doi.org/10.1109/TIP.2022.3169693
34. Y. Cho, W. J. Kim, S. Hong, S. Yoon, Part-based pseudo label refinement for unsupervised person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 7308–7318. https://doi.org/10.1109/CVPR52688.2022.00716
35. S. T. He, H. Luo, P. C. Wang, F. Wang, H. Li, W. Jiang, Transreid: Transformer-based object re-identification, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 15013–15022. https://doi.org/10.1109/ICCV48922.2021.01474
36. T. L. Chen, S. J. Ding, J. Y. Xie, Y. Yuan, W. Chen, Y. Yang, et al., Abd-net: Attentive but diverse person re-identification, in *IEEE/CVF International Conference on Computer Vision (ICCV)*, (2019), 8351–8361. https://doi.org/10.1109/ICCV.2019.00844
37. Y. J. Ge, F. Zhu, D. P. Chen, R. Zhao, H. Li, Self-paced contrastive learning with hybrid memory for domain adaptive object re-id, *arXiv preprint*, (2020), arXiv:2006.02713. https://doi.org/10.48550/arXiv.2006.02713
38. Z. D. Wang, J. W. Zhang, L. Zheng, Y. Liu, Y. Sun, Y. Li, et al., Cycas: Self-supervised cycle association for learning re-identifiable descriptions, in *Computer Vision–ECCV 2020*, Springer, 2020. https://doi.org/10.1007/978-3-030-58621-8_5
39. Y. X. Ge, D. P. Chen, H. S. Li, Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification, *arXiv preprint*, (2020), arXiv:2001.01526. https://doi.org/10.48550/arXiv.2001.01526
40. H. Chen, Y. H. Wang, B. Lagadec, A. Dantcheva, F. Bremond, Joint generative and contrastive learning for unsupervised person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2021), 2004–2013. https://doi.org/10.1109/CVPR46437.2021.00204
41. M. J. Wang, B. S. Lai, J. Q. Huang, X. Gong, X. Hua, Camera-aware proxies for unsupervised person re-identification, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **35** (2021), 2764–2772. https://doi.org/10.1609/aaai.v35i4.16381
42. X. Y. Zhang, D. D. Li, Z. G. Wang, J. Wang, E. Ding, J. Shi, et al., Implicit sample extension for unsupervised person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2022), 7359–7368. https://doi.org/10.1109/CVPR52688.2022.00722
43. L. X. He, J. Liang, H. Q. Li, Z. Sun, Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2018), 7073–7082. https://doi.org/10.1109/CVPR.2018.00739
44. H. Luo, W. Jiang, X. Fan, C. Zhang, Stnreid: Deep convolutional networks with pairwise spatial transformer networks for partial person re-identification, *IEEE Trans. Multimedia*, **22** (2020), 2905–2913. https://doi.org/10.1109/TMM.2020.2965491
45. W. S. Zheng, X. Li, T. Xiang, S. Liao, J. Lai, S. Gong, Partial person re-identification, in *IEEE International Conference on Computer Vision (ICCV)*, (2015), 4678–4686. https://doi.org/10.1109/ICCV.2015.531

46. S. R. Zhou, J. Wu, F. Zhang, P. Sehdev, Depth occlusion perception feature analysis for person re-identification, *Pattern Recognit. Lett.*, **138** (2020), 617–623. https://doi.org/10.1016/j.patrec.2020.09.009
47. Y. F. Sun, Q. Xu, Y. L. Li, C. Zhang, Y. Li, S. Wang, et al., Perceive where to focus: Learning visibility-aware part-level features for partial person re-identification, in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, (2019), 393–402. https://doi.org/10.1109/CVPR.2019.00048

AIMS Press