Signal, Image and Video Processing (2022) 16:1691–1699
https://doi.org/10.1007/s11760-021-02125-8

ORIGINAL PAPER

Pedestrian reidentification based on multiscale convolution feature fusion

Kaiyang Liao[1,3] · Gang Huang[1] · Yuanlin Zheng[1,2] · Guangfeng Lin[1] · Congjun Cao[1,3]

Received: 15 July 2021 / Revised: 18 December 2021 / Accepted: 19 December 2021 / Published online: 12 January 2022

Abstract

The current pedestrian reidentification method based on convolutional neural networks still cannot solve the problems of pedestrian posture change, occlusion and background clutter. Many people use local feature learning or global feature learning alone to alleviate this problem, but they ignore their relevance. Aiming at the difference in emphasis between local features and global features, we propose a unified fusion algorithm, which inherits their advantages while discarding their shortcomings. While random erasure is used to enhance the robustness of the network model, the combined optimization function is used to optimize features of different scales, and the features processed at different scales are merged and spliced to obtain the final representation. Finally, multiple optimization reordering strategies are used to improve the performance of the algorithm. The proposed fusion algorithm was tested on three public pedestrian reidentification datasets, which proved the effectiveness of the method.

Keywords Pedestrian reidentification · Multiscale convolution feature · Deep learning

1 Introduction

Pedestrian reidentification is a process of judging whether a pedestrian is the same target through multiple camera views. It has been widely used in video analysis and pedestrian retrieval for tracking tasks. However, in real life, pedestrian reidentification is affected by factors such as angle of view, illumination, posture, background clutter and occlusion, which makes the difference of pedestrian images in nonoverlapping camera views large. Reducing the impact of this difference on pedestrian reidentification is a huge problem and a severe challenge in current pedestrian reidentification.

The deep learning currently provides a powerful adaptive method to deal with computer vision problems without too much manual manipulation of images and is widely used in the field of pedestrian reidentification. Part of the research focuses on learning features and metrics through a convolutional neural network (CNN) framework, recoding pedestrians as a sorting task and inputting image pairs [1] or triples [2] into a CNN. Because deep learning relies on a large number of sample labels, this method [3] has limitations in the field of pedestrian reidentification.

Deep convolutional neural networks [9] have proven the breakthrough accuracy of pedestrian reidentification, and a series of feature extractors learned from CNNs have been used in other computer vision tasks. Different levels of features have their own advantages. Low-level features [11] have higher resolution and contain more position and detailed information, which can be used to measure fine-grained similarity. However, due to the lower number of convolutional layers it passes through, it contains more noise. The semantics are not strong, and they are easily affected by background confusion and semantic clutter. High-level features [12] have stronger semantic information, which is used to measure semantic similarity, but their resolution is low, their ability to

✉ Kaiyang Liao
liaokaiyang@xaut.edu.cn

Gang Huang
1106160413@qq.com

Yuanlin Zheng
zhengyuanlin@xaut.edu.cn

[1] College of Faculty of Printing, Packaging Engineering and Digital Media Technology, Xi'an University of Technology, Xi'an 710048, China

[2] Key Lab of Printing and Packaging Engineering of Shaanxi Province, Xi'an 710048, China

[3] Printing and Packaging Engineering Technology Research Centre of Shaanxi Province, Xi'an 710048, China

perceive details is poor, and they are not able to describe the fine-grained details of the image. Therefore, how to effectively combine the two is the key to improving recognition accuracy. This paper extracts and encodes convolution features from different levels, stitches these different levels of convolution features to test images and uses the complementarity of low-level and high-level features to improve the similarity measurement between the query image and other candidate images.

To relieve the pressure of reidentification tasks caused by complex background or pedestrian posture changes and learn the global information of pedestrians and effective local discriminant features, this paper proposes a multiscale learning experimental design based on deep feature fusion.

With ResNet50 [17] as the basic framework, a multilearning branch network structure is designed, including global feature fusion and local feature fusion. First, the global fusion learning branch captures the approximate attention to pedestrians from the entire image and learns the multilevel feature information of pedestrians. Second, the local fusion learning branch extracts local features from different local areas and learns the deep-level local features of pedestrians. The network pays more attention to the correlation between features to learn more distinctive features and provide more representative and spatially distributed features by fusing the features of the four branches.

This paper uses the complementary advantages of different levels of convolutional features to propose a pedestrian reidentification method model based on multiscale convolutional feature fusion. The proposed multiscale convolution feature fusion method is shown in Fig. 1. In this paper, ResNet50 is selected as the backbone network, and a multibranch joint learning experimental network including global feature fusion and local feature fusion is designed. The global feature learning branch captures the most significant information among all different pedestrians and learns recognizable features; the local feature fusion learning branch supplements the global features to learn more fine-grained features. This strengthens the learning of the correlation between the nonadjacent parts of pedestrians and makes the network pay more attention to the correlation between features. By fusing the characteristics of the four branches, it provides more representative and spatially distributed characteristics.

The following four parts are important:

1. Making full use of the shallow information and high-dimensional semantic information of the image to fuse multiscale features to achieve information complementation, and the recognition accuracy is mentioned;
2. Using the random erasure data enhancement method and dynamic learning rate adjustment method to enhance the robustness of the network model;
3. Using the combined optimization loss function, by combining the superior performance of multiple loss functions, resulting in the model being optimized to improve the accuracy of the classifier;
4. Adopting the reordering strategy of multimethod optimization, and multidistance optimization is used to make the matching result obtain a higher ranking.

2 Pedestrian reidentification method model based on multiscale convolution feature fusion

This paper proposes a new pedestrian reidentification algorithm based on the principle of multiscale convolution feature fusion to improve the accuracy of pedestrian reidentification. The backbone network in this article uses the ResNet-50 network. Specifically, the step size of the fourth stage of ResNet-50 is set from 2 to 1, and the size of the convolution feature map extracted through the backbone network becomes 1/16 of the original input image size. When the input image size is 256×128, after the second stage of ResNet-50, a feature map with a spatial size of 8×4 will be output. After setting the stride from 2 to 1, a feature map with a size of 16×8 can be obtained. This operation does not involve additional training parameters. Increasing the size of the feature map also improves the spatial resolution. In the ResNet-50 backbone network, using random erasure, warm-up learning rate setting and other training skills to constantly optimize the network model ensures that pedestrian reidentification can obtain a higher recognition rate.

2.1 Optimized network model

2.1.1 Random erasure strategy

At this time, data enhancement operations need to be performed on the pictures. Random erasing augmentation (REA) [16] is a data extension method of random erasure. The basic idea is to randomly select a block in the image to cover up the noise block. Random erasure is a method of data expansion that can reduce the degree of model overfitting, so it can improve the performance of the model.

In the training process, random erasure has a certain probability. For the image I in the minibatch batch, suppose the probability of it being randomly erased is P, and the probability of keeping the whole unchanged is $1-P$. In the process of data preprocessing, different training images will be generated.

A rectangular area I_e in the original image is randomly selected, then pixels in the rectangular area are erased and

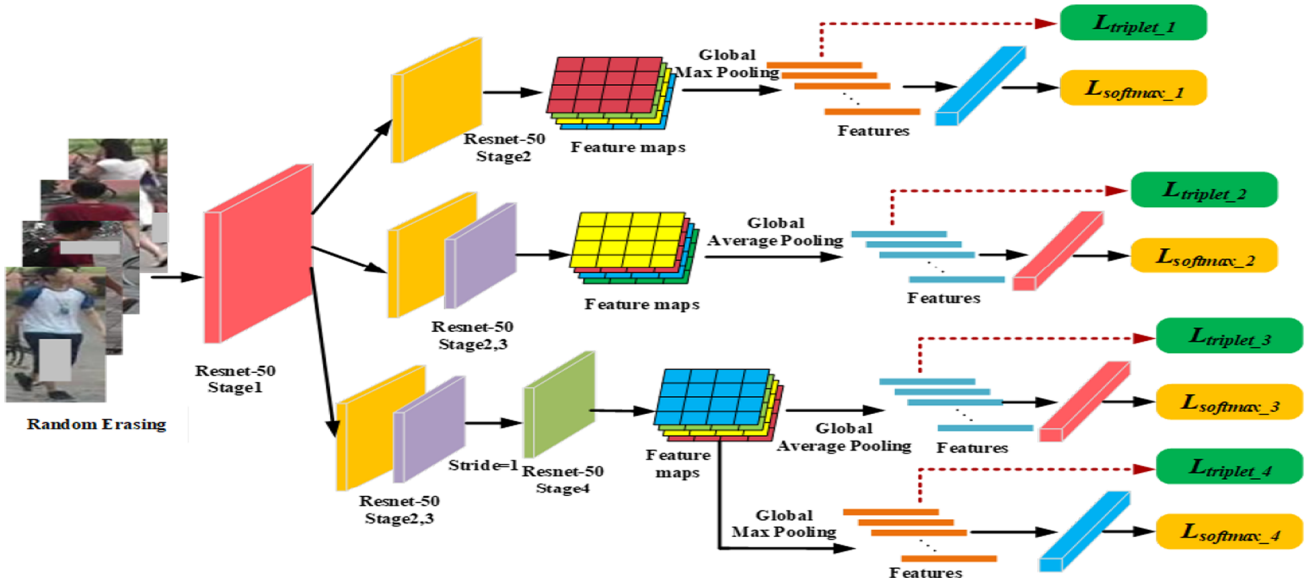

Fig. 1 Person reidentification flow chart based on multiscale convolution feature fusion

replaced with random values. Suppose the area of the image that needs to be input to the network model for training is:

$$S = W \times H \tag{1}$$

where W represents the width of the pedestrian image, and H represents the height of the pedestrian image.

The area value of the erased area in the original image is randomly initialized to S_e, where S_e/S is within the range specified by the minimum value S_l and the maximum value S_h. The aspect ratio of the erased area is initialized randomly between r_1 and r_2 and set to r_e. The size of I_e is:

$$H_e = \sqrt{S_e \times r_e} \tag{2}$$

$$W_e = \sqrt{S_e / r_e} \tag{3}$$

where S_e represents the area value of the erased rectangular frame; r_e is the aspect ratio of the erased rectangular frame; H_e is the height of the erased rectangular frame, and W_e is the width of the erased rectangular frame.

A point $P = (x_e, y_e)$ is randomly initialized in the pedestrian image I if the following conditions are met:

$$x_e + W_e \leq W \tag{4}$$

$$y_e + H_e \leq H \tag{5}$$

Then $(x_e, y_e, x_e + W_e, y_e + H_e)$, is used as the coordinate value of the selected rectangular area. If the above conditions are not met, the above process will be repeated until an appropriate I_e is selected. Using the selected erase area I_e, each pixel in the rectangular frame I_e is assigned to a random value in the [0, 255] range. Finally, the randomly erased pictures are output, and the result of the pictures after random erasure processing is shown in Fig. 2.

2.1.2 Combinatorial optimization loss function

In deep learning, the algorithm-solving process is actually the process of solving or optimizing the objective function by designing the corresponding algorithm. Different loss functions have different focuses. Therefore, we propose a combined optimization loss function to optimize the network. By combining the advantages of different loss functions, the performance of the classifier is improved. This section introduces the cross entropy loss function and triple loss function we use in the network.

The cross-entropy loss function (softmax loss) is widely used in various multiclassification tasks. The formula of the softmax function is of this form:

$$S_i = e^{z_i} \Big/ \sum_k e^{z_k} \tag{6}$$

where S_i is the output of the i-th neuron.

And the output z_i of the neuron is set as:

$$z_i = \sum_j w_{ij} x_{ij} + b \tag{7}$$

where w_{ij} represents the j-th weight of the i-th neuron; x_{ij} is the input neuron; b is the bias value of each neuron, and z_i is the i-th output of the network.

Fig. 2 Random erasing preprocessing effect diagram

When a softmax function is added to this output, it becomes:

$$a_i = e^{z_i} \Big/ \sum\nolimits_k e^{z_k} \tag{8}$$

a_i represents the size of the probability value of class i corresponding to this input image. The value range of each type of a_i is in the interval [0, 1]. After the probability values of all the categories are obtained, the softmax function is added behind the neural network. Therefore, the loss function of softmax is:

$$L_{\text{soft max}} = \sum -\hat{y}_i \ln y_i \tag{9}$$

where y_i indicates that the output of the neuron can also be used as the prediction result; $\hat{y}_i$ is the true value of the i-th category, and $\hat{y}_i$ can only take the value 0 or 1.

Generally, in pedestrian reidentification research, only the combination of ResNet-50 and the softmax loss function is used as the backbone network, which achieves good results on large datasets. However, the model using only the softmax loss function lacks the ability to distinguish the fine-grained features of pedestrian images.

In the field of pedestrian reidentification, triplet loss is also widely used and more often combined with softmax loss in the network model. The formula for calculating the loss function after feature extraction is as follows:

$$L_{\text{triplet}} = \sum_{i}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+ \tag{10}$$

where $\left\| f(x_i^a) - f(x_i^p) \right\|_2^2$ represents the Euclidean distance measurement value of the positive sample and the anchor point sample, which is the distance within the class; $\left\| f(x_i^a) - f(x_i^n) \right\|_2^2$ is the Euclidean measurement value of the negative sample and the anchor point sample, which represents the distance between the classes; α is the distance between x_i^a and x_i^n, and there is a minimum interval between x_i^a and x_i^p.

Through the triplet loss function, the network model can shorten the distance between pedestrian images with the same label and extend the distance between pedestrian images with different labels, making the trained network model more discriminative. A schematic diagram of the triple loss is shown in Fig. 3.

This article uses four softmax losses and four triplet losses. The final loss function is expressed as:

$$L = \frac{1}{m} \left(\sum_{1}^{m} L_{\text{soft max}} + \sum_{1}^{m} L_{\text{triplet}} \right) \tag{11}$$

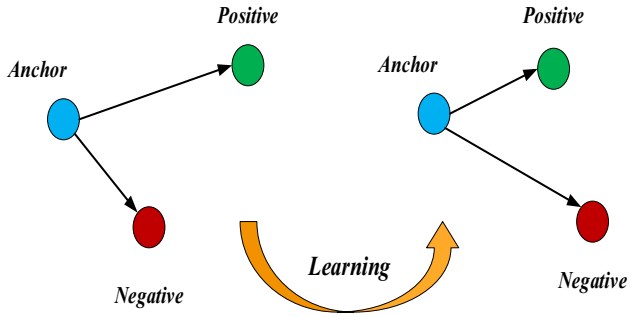

Fig. 3 Triplet loss

where m represents the number of loss functions, which is set to 4 in this article.

2.1.3 Dynamic learning rate

The learning rate has a great influence on the performance of the pedestrian reidentification model. This article uses the simplest linear strategy, that is, the first 10 epochs of learning gradually increase from 0 to the initial learning rate. In practice, the first 10 cycles are used to increase the learning rate linearly from 3.5×10^{-5} to 3.5×10^{-4}. Then, in the 40th and 70th learning cycles, the learning rate drops to 3.5×10^{-5} and 3.5×10^{-6}, respectively. The learning rate $l_r(t)$ in the t-th period is calculated as:

$$l_r(t) = \begin{cases} 3.5 \times 10^{-5} \times \frac{t}{10} & \text{if } t \leq 10 \\ 3.5 \times 10^{-4} & \text{if } 10 < t \leq 40 \\ 3.5 \times 10^{-5} & \text{if } 40 < t \leq 70 \\ 3.5 \times 10^{-6} & \text{if } 70 < t \leq 120 \end{cases} \tag{12}$$

2.1.4 Training process

The network model in this paper is trained using 2 GTX2080Ti GPUs (batch size is 32). Each pedestrian identity includes 4 images, so there are 8 pedestrian identities in each batch. The backbone network ResNet-50 is initialized using ImageNet pretraining. The SGD optimizer is used to optimize the model, and the combination of softmax loss and triplet loss is used to continuously optimize the network, making the model more robust.

2.2 Feature extraction based on Resnet-50 neural network

2.2.1 Pooling strategy

In this paper, a global average pooling (GAP) layer [21] is used, which is used to replace the fully connected layer in the CNN. Specifically, as shown in Fig. 4, the feature map can be easily interpreted as a category confidence map. The advantage of the global average strategy compared to the fully connected layer is that the convolution structure can be retained better by enhancing the correspondence between the feature map and the category, and the global average pooling layer has no parameter settings, which avoids overfitting in this layer.

Similar to global average pooling, this paper also uses global max pooling (GMP) operations to perform global maximum pooling for the feature maps obtained at different stages because global maximum pooling encourages the network to identify relatively weak image salient features. During the test, the features obtained by global maximum pooling and global average pooling are stitched together as the embedding vector of the pedestrian image.

2.2.2 Multiscale convolution feature extraction

When training the network model, the global average pooling and global maximum pooling strategies are used to pool the feature maps to obtain multiscale local feature vectors. The pooled feature vector is used in the calculation of the triple loss function; the different feature vectors obtained are classified; the normalized weight is added to each feature, and the classification softmax loss function is used to improve the classification performance. Finally, the gradient descent method is applied to the network model. Specifically, the ResNet-50 network is used as the backbone network. Based on the optimization techniques, the feature maps obtained in the second and third stages of convolution are input into the global maximum pooling layer and the global average pooling layer, respectively. The 1024-dimensional and 2048-dimensional feature vectors containing pedestrian discriminative information are obtained. Then, through a 1×1 convolutional layer, a batch normalization layer and a ReLU layer, the dimension is reduced to 512. After the 4th stage of ResNet-50, the step size of the convolution kernel is changed from 2 to 1 so that the size of the convolutional feature map obtained after the 4th stage of the network model becomes larger. Then, the convolutional feature map containing more pedestrian information is deep-copied into two copies, which are input to the global average pooling layer and the global maximum pooling layer, and then the 1×1 convolution kernel is used to reduce the dimensions of the two pooled feature vectors to 512 dimensions. In the recognition stage, the four 512-dimensional feature vectors obtained by splicing are finally obtained as a new feature vector of 2048 dimensions, and multiple different feature vectors are merged to obtain the recognition result.

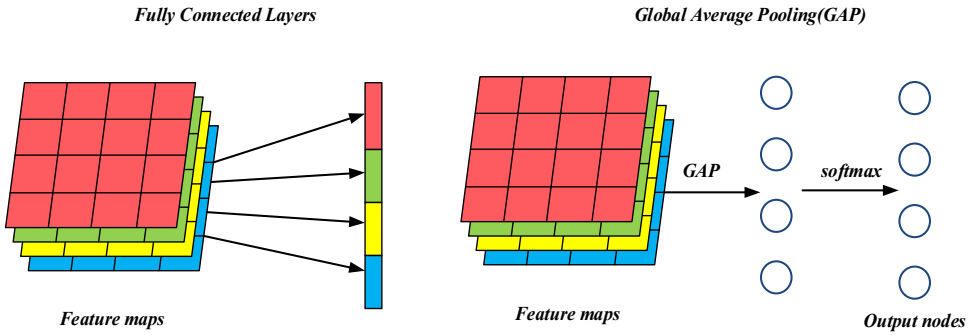

Fig. 4 Global average pooling diagram

2.3 Reordering strategy of multiple optimization methods

The main advantage of many reranking methods [14] is that they can be implemented without additional training samples and can be applied to any initial ranking results. We propose a combined distance optimization reordering strategy. Through the combined use of the Mahalanobis distance and Jacobian distance, the result is closer to the expectation.

Zhong et al. [15] proposed a k-reciprocal coding method to reorder the results of pedestrian reidentification. Specifically, given a query image, the k-reciprocal feature can be calculated by encoding its k-reciprocal nearest neighbor as a single vector, which is used to reorder under the Jaccard distance, and the final distance is calculated as the original combination of distance and Jacobian distance.

Given a pedestrian p in a test image and a set of reference images $G = \{g_i | i = 1, 2, \ldots, N\}$, the original distance between the two pedestrian images p and g_i can be measured by the Mahalanobis distance,

$$d(p, g_i) = \left(x_p - x_{g_i}\right)^{\mathrm{T}} M \left(x_p - x_{g_i}\right) \tag{13}$$

where x_p represents the appearance feature of test image p; x_{g_i} is the appearance feature of reference image g_i, and M is a positive semidefinite matrix.

The initial ranking list is obtained according to the original distance between the test image P and the reference image g_i:

$$L(p, G) = \left\{g_1^0, g_2^0, \ldots, g_N^0\right\} \tag{14}$$

The goal is to resort $L(p, G)$ so that more correctly matched pedestrian samples appear in the front row of the sorted list to improve the reidentification performance.

The first k samples in the sorted list are defined, namely, k-nearest neighbors (k-nn):

$$N(p, k) = \left\{g_1^0, g_2^0, \ldots, g_k^0\right\}, \quad |N(p, k)| = k \tag{15}$$

k-reciprocal nearest neighbors (k-rnn), expressed as:

$$R(p, k) = \{g_i | (g_i \in N(p, k)) \wedge p \in N(g_i, k)\}$$

To solve the problem that the matched sample is not in the k-nearest neighbor sample due to a series of changes in illumination, viewing angle and pedestrian posture, a more robust set is defined:

$$\begin{aligned} &R^*(p, k) \leftarrow R(p, k) \cup R\left(q, \frac{1}{2}k\right) \\ &\text{s.t.} \left| R(p, k) \cap R\left(q, \frac{1}{2}k\right) \right| \geq \frac{2}{3} \left| R(q, \frac{1}{2}k) \right|, \quad \forall q \in R(p, k) \end{aligned} \tag{16}$$

For each test sample q in the original set $R(p, k)$, their k-reciprocal nearest neighbor set $R(q, \frac{1}{2}k)$ is found. When the number of coincidence samples reaches a certain condition, its union with $R(p, k)$ is found, and the positive samples that are not matched in the original set $R(p, k)$ are reincluded.

To redistribute the weight to each element according to the original distance, a Gaussian kernel is used to encode the k-reciprocal nearest neighbor set of the retrieved image into an N-dimensional vector, which is defined as $v_p = \left[v_{p,g_1}, v_{p,g_2}, \ldots, v_{p,g_N}\right]$, and v_{p,g_i} set as:

$$v_{p,g_i} = \begin{cases} e^{-d(p,g_i)} & \text{if } g_i \in R^*(p, k) \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

The cardinality of the intersection and union used when calculating the Jacobian distance is rewritten as:

$$\left| R^*(p, k) \cap R^*(g_i, k) \right| = \left\| \min\left(v_p, v_{g_i}\right) \right\|_1 \tag{18}$$

$$\left| R^*(p, k) \cup R^*(g_i, k) \right| = \left\| \max\left(v_p, v_{g_i}\right) \right\|_1 \tag{19}$$

The intersection takes the smallest value in the corresponding dimension of the two feature vectors as the degree to which they both contain g_i through the minimum operation. The biggest operation of union is to count the total set of matching candidates in the two sets.

The final Jacobian distance is as follows:

$$d_J(p, g_i) = 1 - \frac{\sum_{j=1}^{N} \min\left(v_{p,g_j}, v_{g_i,g_j}\right)}{\sum_{j=1}^{N} \max\left(v_{p,g_i}, v_{g_i,g_j}\right)} \tag{20}$$

The final calculated distance is as follows:

$$d^*(p, g_i) = (1 - \lambda) d_J(p, g_i) + \lambda d(p, g_i) \tag{21}$$

The initial ranking is reranked by combining the original Mahalanobis distance and Jacobian distance. The final distance is the weighted sum of the two distances. The weighting parameter is mainly used to measure the relative importance of the two distances. $\lambda = 0.3$ is set in the experiments.

3 Experimental results and analysis

In this section, to verify the effectiveness of the multiscale convolution feature fusion algorithm in this paper, three commonly used pedestrian reidentification datasets are tested, including the Market-1501[18], CUHK03 [19] and DukeMTMC-reID [20] datasets. It also follows the latest strategies to generate training, query and gallery data. The original CUHK03 dataset is divided into 20 random training/testing groups for cross-validation, which is usually used in methods based on manual functions. The new partition method used in the experiment further separates the training image from the candidate image and selects the challenging query image for evaluation. Therefore, data integration using CUHK03 dataset is the most challenging task.

We evaluate the impact of random erasure probability P on the model in the CHUK03 database. When the parameters of other data enhancement methods are fixed, the image size is 256×128, and the minimum aspect ratio of the deleted area is fixed. As shown in Fig. 5, when the probability of random erasure is $P = 0.5$, the model obtains the best performance.

We also study the influence of different image sizes on the pedestrian reidentification model. On the basis of adding optimization techniques such as random erasure data enhancement, dynamic learning rate mechanism and stride change, we set the number of batch trainings to 32 and the probability of random erasure to 0.5 and set the image size to 256×128, 224×224, 384×128 and 384×192. Four different models are trained on the Market-1501 and DukeMTMC-reID datasets. The experimental results are shown in Table 1. The four models show similar performance on the two datasets. The performance of the image size of 256×128 is better, so this article uses the image size of $256 \times 128$to train the model.

The multiscale convolution feature fusion model designed in the article is used as a comparison model. Without any

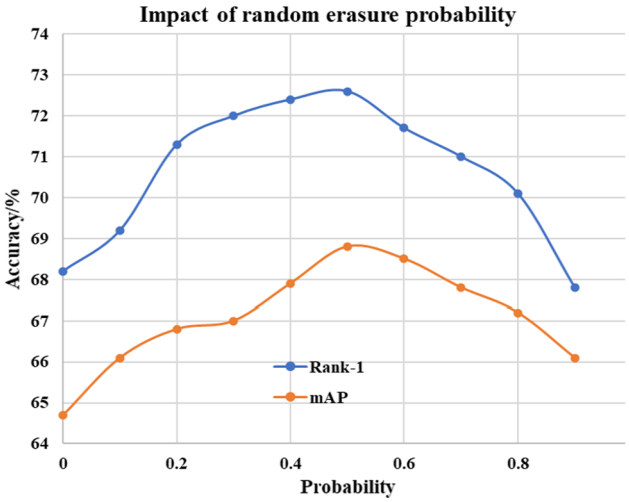

Fig. 5 Impact of random erasure probability

Table 1 Performance of ReID models with different image sizes

Image size	Market-1501		DukeMTMC-reID	
	Rank-1	mAP	Rank-1	mAP
256×128	96.0	87.3	89.7	80.1
224×224	95.4	86.5	89.3	79.3
384×128	95.9	87.4	87.4	78.8
384×192	95.8	87.2	87.6	78.9

training skills, the Rank-1 values of the proposed model on Market1501 and DukeMTMC-reID reach 90.7% and 82.2%, respectively. The dynamic learning rate mechanism, random erasure enhancement and stride changes are added to the model training process one by one. Finally, these techniques enabled the benchmark to obtain a Rank1 value of 96.0% and a mAP value of 87.3% on Market1501. On DukeMTMC-reID, the Rank1 value is 89.7%, and the mAP value is 80.1%.

Table 2 shows the statistical comparison of the performance of the multiscale convolutional feature fusion network in this paper and the latest method on the CUHK03, DukeMTMC-reID and Market-1501 datasets. This article is related to IDE [4], PAN [22], SVDNet [23], DaRe [24], HA-CNN [25], PCB [10], PCB + RPP [10], BDB [13] and MGN [26], and other similar methods are compared. The algorithm model proposed in this paper has achieved better performance in the three major databases, and the greatest improvement has been achieved with the most challenging dataset CUHK03. Compared with the BDB method in the Lable database, the Rank1 value is 11% higher and the mAP value is nearly 12% higher. For the DukeMTMC-reID database, compared with the MGN method, the Rank1 value of the model proposed in this paper is nearly 3% higher, and the mAP value is 10% higher. Compared with these mainstream algorithms, the data in Table 2 prove that the

Table 2 Performance comparison of our algorithm with state-of-the-art algorithms

Method	Market-1501		CUHK03-Lable		CUHK03-Detect		DukeMTMC-reID	
	Rank − 1	mAP	Rank − 1	mAP	Rank − 1	mAP	Rank − 1	mAP
IDE	72.5	46.0	22.2	21.0	21.3	19.7	67.7	47.1
PAN	82.8	63.4	36.9	35.0	36.3	34.3	71.6	51.5
SVDNet	82.3	62.1	–	–	41.5	37.3	76.7	56.8
DaRe	89.0	76.0	66.1	61.6	63.3	59.0	80.2	64.5
HA-CNN	91.2	75.7	44.4	41.0	41.7	38.6	80.5	63.8
PCB	92.4	77.3	–	–	61.3	54.2	81.9	65.3
PCB + RPP	93.8	81.6	–	–	62.8	56.7	83.3	69.2
BDB	94.2	84.3	73.6	71.7	72.8	69.3	86.8	72.1
MGN	95.7	86.9	68.0	67.4	66.8	66.0	88.7	78.4
Proposed	96.0	87.3	73.8	73.8	73.5	71.2	89.7	80.1
Proposed + rerank	96.8	94.1	84.5	83.6	81.9	79.2	91.3	88.9

multiscale convolution feature fusion method proposed in this paper is effective. This paper also uses the reordering strategy. Compared with not using the reordering method, in the three major databases, the Rank1 value and mAP value were improved. On the CUHK03 dataset, the performance improvement obtained by using the reordering method is the most significant and further proves the effectiveness of the reordering algorithm.

4 Conclusion

This paper proposes a pedestrian rerecognition model based on multiscale convolution feature fusion, the model consists of different levels of convolutional features, and uses the complementary features of low-level features and high-level features. First, using Resnet-50 as the backbone network, the model is made more robust through optimization methods such as stride change, dynamic learning rate mechanism and random erasure enhancement. Second, through multiple branch networks such as global fusion and local feature fusion, different levels of representation information are learned. Based on the combination of low-level and high-level features, the low-level information and high-level semantic information of the image are fully utilized to collect spatial information and avoid models. By fusing features of different scales, the learned features are more representative. In addition, it also combines multiple loss functions to supervise the training network, which improves the generalization ability of the model. The results on the Market-1501, DukeMTMC-reID and CUHK03 datasets verify the effectiveness of the algorithm model proposed in this paper.

Acknowledgements This work is supported by the National Natural Science Foundation of China Project Nos. 61771386, 52075435 and Natural Science Foundation of Shaanxi Province No. 2021JM-340.

References

1. Wu, S., et al.: An enhanced deep feature representation for person re-identification. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016)
2. Ding, S., et al.: Deep feature learning with relative distance comparison for person re-identification. Pattern Recognit. **48**(10), 2993–3003 (2015)
3. Lu, X., Yuan, Y., Zheng, X.: Joint dictionary learning for multispectral change detection. IEEE Trans. Cybern. **47**(4), 884–897 (2016)
4. Zheng, L., et al.: Person re-identification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
5. Tao, D., et al.: Deep multi-view feature learning for person re-identification. IEEE Trans. Circuits Syst. Video Technol. **28**(10), 2657–2666 (2017)
6. Kumar, V., et al.: Pose-aware person recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
7. Su, C., et al.: Pose-driven deep convolutional model for person re-identification. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
8. Suh, Y., et al.: Part-aligned bilinear representations for person re-identification. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
9. Cheng, D., et al.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
10. Sun, Y., et al.: Beyond part models: person retrieval with refined part pooling (and a strong convolutional baseline). In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
11. Xiao, T., et al.: Joint detection and identification feature learning for person search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
12. Ahmed, E., Jones, M., Marks T.K.: An improved deep learning architecture for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2015)
13. Dai, Z., et al.: Batch dropblock network for person re-identification and beyond. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2019)

14. Bai, S., Bai, X., Tian, Q. Scalable person re-identification on supervised smoothed manifold. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
15. Zhong, Z., et al.: Re-ranking person re-identification with k-reciprocal encoding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017)
16. Zhong, Z., et al.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34(07) (2020)
17. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016)
18. Zheng, L., et al.: Scalable person re-identification: a benchmark. In: Proceedings of the IEEE International Conference on Computer Vision (2015)
19. Li, W., et al.: Deepreid: deep filter pairing neural network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2014)
20. Ristani, E., et al.: Performance measures and a data set for multi-target, multi-camera tracking. In: European Conference on Computer Vision. Springer, Cham (2016)
21. Lin, M., Chen, Q., Yan, S.: Network in network (2013). arXiv preprint https://arxiv.org/1312.4400
22. Zheng, Z., Zheng, L., Yang, Y.: Pedestrian alignment network for large-scale person re-identification. IEEE Trans. Circuits Syst. Video Technol. **29**(10), 3037–3045 (2018)
23. Sun, Y., et al.: Svdnet for pedestrian retrieval. In: Proceedings of the IEEE International Conference on Computer Vision (2017)
24. Wang, Y., et al.: Resource aware person re-identification across multiple resolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
25. Li, W., Zhu, X., Gong, S.: Harmonious attention network for person re-identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018)
26. Wang, G., et al.: Learning discriminative features with multiple granularities for person re-identification. In: Proceedings of the 26th ACM International Conference on Multimedia (2018)