Hindawi
Mathematical Problems in Engineering
Volume 2020, Article ID 3731848, 11 pages
https://doi.org/10.1155/2020/3731848

Research Article

A Person Reidentification Algorithm Based on Improved Siamese Network and Hard Sample

Guangcai Wang,[1] **Shiqi Wang**,[1] **Wanda Chi**,[1] **Shicai Liu**,[2] **and Di Fan**[1]

[1]*College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao 266590, China*
[2]*College of Intelligent Equipment, Shandong University of Science and Technology, Tai'an 271000, China*

Correspondence should be addressed to Di Fan; skd992372@sdust.edu.cn

Received 29 April 2020; Accepted 2 June 2020; Published 24 June 2020

Guest Editor: Weicun Zhang

Person reidentification is aimed at solving the problem of matching and identifying people under the scene of cross cameras. However, due to the complicated changes of different surveillance scenes, the error rate of person reidentification exists greatly. In order to solve this problem and improve the accuracy of person reidentification, a new method is proposed, which is integrated by attention mechanism, hard sample acceleration, and similarity optimization. First, the bilinear channel fusion attention mechanism is introduced to improve the bottleneck of ResNet50 and fine-grained information in the way of multireceptive field feature channel fusion is fully learnt, which enhances the robustness of pedestrian features. Meanwhile, a hard sample selection mechanism is designed on the basis of the P2G optimization model, which can simplify and accelerate picking out hard samples. The hard samples are used as the objects of similarity optimization to realize the compression of the model and the enhancement of the generalization ability. Finally, a local and global feature similarity fusion module is designed, in which the weights of each part are learned through the training process, and the importance of key parts is automatically perceived. Experimental results on Market-1501 and CUHK03 datasets show that, compared with existing methods, the algorithm in this paper can effectively improve the accuracy of person reidentification.

1. Introduction

As an important intelligent video analysis technology, person reidentification is widely used in the fields of intelligent security, case detection, lost query, intelligent interaction, and so on. It is an important means in the field of public safety. However, the huge difference in lighting, occlusion, resolution, background, and human posture, in the actual scene, and the deficiency of data make person reidentification task still face many difficulties and challenges.

Feature learning method and metric learning method are two basic directions in the research of person reidentification task. The method of feature representation mainly focuses on how to extract robust person features, such as the Ensemble of Localized Features [1], like colour and texture features, SDALF (Symmetry-Driven Accumulation of Local Features) [2], the Local Maximal Occurrence (LOMO) representation [3], and the fusion net of CNN features and manual features (FFN) [4]. The weighted histogram feature of middle dense and significant blocks is based on the Gauss mixture model [5]. Using the information continuity between image blocks, we design the attention mechanism to optimize the image similarity of blocks after the PCB network [6]. However, only relying on image feature expression cannot completely solve this task. Many scholars have done a lot of research in image feature measurement. KISSME algorithm [7] measures the degree of difference between samples by likelihood ratio and obtains the Mahalanobis metric, which reflects the log-likelihood ratio property. Based on the Siamese Network, using contrastive loss [8–11] and verification loss, like Triplet loss [12–16] and Quadruplet loss [17], makes the distance between the same ID persons closer and that between those with different ID farther. Some rerank methods, like SSM (Supervised Smoothed Manifold)

[18], *k*-reciprocal neighbours [19] and UED (the Unified Ensemble Diffusion) [20], were used to achieve more accurate person reidentification.

Aiming at extracting more robust feature for person images, based on the Siamese Network, we improve the convolution settings on the ResNet50, integrate multiscale visual field features, and design the attention mechanism of channel fusion to enhance the expression of fine-grained feature information in the image. At the same time, unlike the existing feature measurement and rerank algorithms, based on the P2G similarity optimization model [21], a hard sample selection mechanism is set up, which uses the hard samples to optimize the P2G similarity to enhance the learning and generalization ability of the model. In addition, in order to make full use of the detailed information of the image features, while avoiding the cumbersome process of local feature extraction, the features are grouped in the way of horizontal overlap, and the similarity between query image and gallery image for each group was calculated, respectively. We integrate all groups of similarities and global similarity to realize automatically perceiving of the key parts, so as to get more accurate similarity measurement results. Experiments on Market-1501 and CUHK03 datasets show that the proposed model can extract person features completely and achieve higher recognition accuracy.

2. Materials and Methods

In this paper, a complete person reidentification model is constructed, in which the improved ResNet50 is the backbone to extract person image features. Based on the P2G similarity optimization model, the hard sample mining and feature grouping and similarities fusion module are introduced to improve the accuracy of person reidentification model. The whole model can be divided into three modules: feature extraction module, hard sample mining module, and feature group and similarities optimization fusion module. Finally, by sorting the similarity scores, we get the retrieval results of the probe in all gallery images. The overall framework of the algorithm is shown in Figure 1.

The feature extraction module adopts the improved ResNet50 as the backbone and designs a bilinear feature channel fusion module to extract more robust features. The hard sample mining module is based on the calculated initial feature distance $d(p, g)$ and selects the most difficult positive samples and negative samples for each probe image as the similarity update examples. On the one hand, it can improve the generalization ability of the network; on the other hand, it can also reduce data redundancy and the computational pressure. The feature grouping module divides the features of the images into 3 groups in the way of horizontal overlap grouping, and each group of features is used to calculate and optimize the local similarity score separately to enhance the impact of detail information on similarity measurement. Finally, combined with global feature similarity, we could obtain a more accurate similarity.

2.1. Improved ResNet Module. For normal convolution, convolution kernels in convolution operations can be regarded as a three-dimensional filter, containing channel dimensions and spatial dimensions (the width and height of feature map), thereby realizing joint mapping of channel correlation and spatial correlation. Convolution neural network reduces parameters in training process through receptive field and weight sharing. However, as the number of network layers increases, the amount of parameters also increases, thus increasing the amount of computation training process and the resource cost.

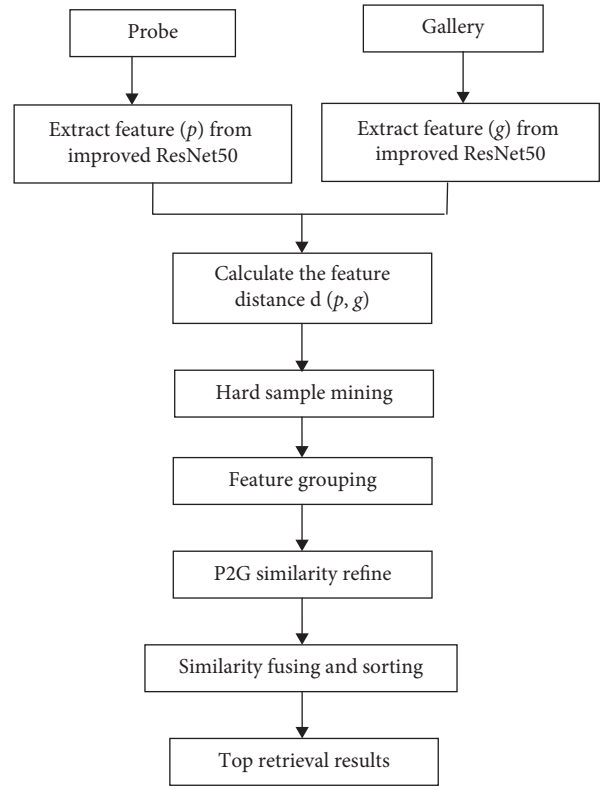

Figure 1: The overall framework flow of the algorithm.

Depthwise separable convolution, using different convolution kernels for different input channels, decomposes the normal convolution into two processes: depthwise convolution (DW) and pointwise convolution (PW). The channel correlation and spatial correlation in convolution layer are separately mapped to realize the decoupling of the two. The depthwise convolution groups features in the channel dimension, and the number of groups is the number of characteristic channels. Compared with normal convolution, the number of convolution kernel parameters and the amount of computation in convolution process are reduced to a great extent, thus reducing the pressure of network computing resources.

Based on the idea of depthwise separable convolution, we improved the residual bottleneck module in the ResNet50 network. First, we extract the channel features and then replace the normal convolution (Conv 3×3) in the original bottleneck by depthwise convolution (DW 3×3) to learn spatial information. Then we use pointwise convolution (PW 1×1) to realize the fusion of channel features. The improved bottleneck block network structure is shown in Figure 2.

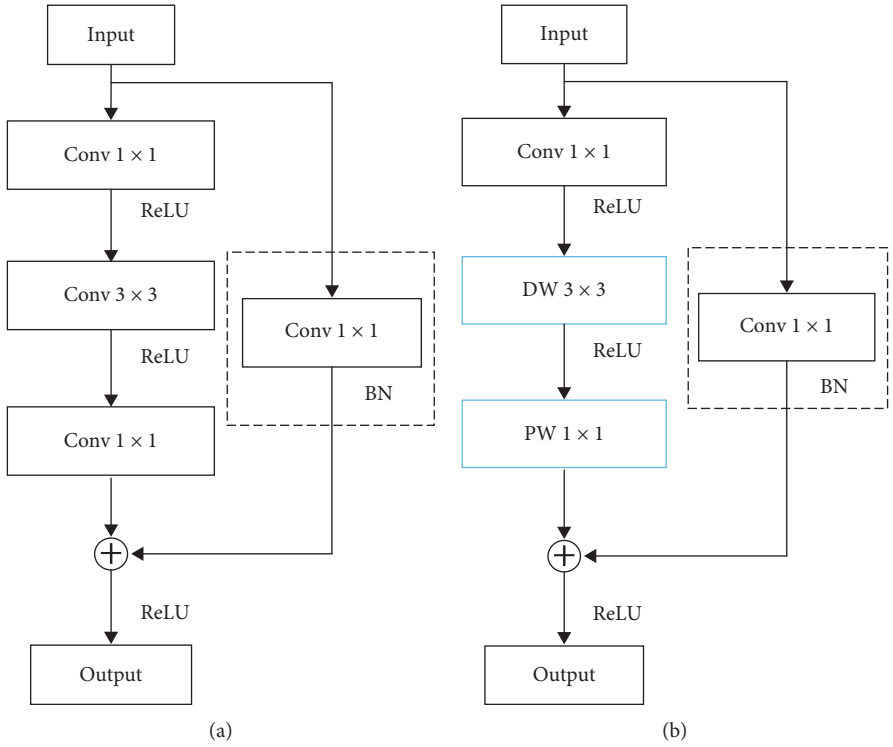

Figure 2: Bottleneck improvement: (a) original bottleneck; (b) improved bottleneck.

The improvement of the bottleneck block realizes the separation of channel features and spatial features, weakens the coupling between channel features and spatial features, and makes the network learn the features of different dimensions more distinguishably. At the same time, independent learning channel features and spatial features strengthen the connection between the spatial structures on the feature map and also make the correlation between different channel features more differentiated. With the deepening of the network, the spatial feature size of the image is getting smaller and smaller, and the number of channels is increasing. We apply the improved bottleneck block to the final stage of the ResNet50 network, without reducing the network's ability to extract features, to minimize the network computing overhead and parameter storage pressure.

Because different channels of image features often contain different attributes or different degrees of expression of the same attribute, the simple addition of features does not make full use of the information of image, and it will waste or even lose the image information to some extent. Bilinear models calculate the outer product of different spatial positions. Moreover, the average convergence of features at different spatial locations has a good effect in learning fine-grained image representation. The outer product captures the pairwise correlation between feature channels and has the characteristic of translation invariance.

A bilinear channel fusion attention mechanism is designed on the basis of the ResNet residual structure with deep separable convolution. On the one hand, interactive model for local pairwise features is carried out, and bilinear expressions are applied in convolution neural networks in depth to learn fine-grained image features (such as clothing texture, hairstyle, and movement of persons). On the other hand, the channel information is broken through, and the attention of different channels is strengthened by means of channel attention. The bilinear channel fusion attention module structure is shown in Figure 3.

In Figure 3, the input feature connects the channel information through a normal convolution (Conv 1×1) to get the initial feature $I \in R^{C \times H \times W}$. We divide the initial feature I into three branches to deal with it separately, and each branch sets different receptive fields and convolution methods, respectively. For Branch 1, the initial feature goes through a layer of depthwise convolution (DW 3×3) to get feature $X_1 \in R^{C \times H \times W}$. For Branch 2, it expands the receptive field of the initial feature through the two layers of depthwise convolution (DW 3×3), which is equivalent to the spatial information that incorporates the scope of the size of 5×5 and gets the feature $X_2 \in R^{C \times H \times W}$. Branch 3 is the original feature I without any processing. Then we vectorize the feature I, as well as each channel of feature X_1 and feature X_2, where $N = H \times W$:

$$\begin{aligned} I \in R^{C\times H\times W} &\xrightarrow{\text{vectorization}} I_1 \in R^{C\times N}, \\ X_1 \in R^{C\times H\times W} &\xrightarrow{\text{vectorization}} F1 \in R^{C\times N}, \\ X_2 \in R^{C\times H\times W} &\xrightarrow{\text{vectorization}} F2 \in R^{C\times N}. \end{aligned} \tag{1}$$

After that, for each location l of the input feature I, we have designed the bilinear model shown in the following equation:

$$\text{bilinear}\left(l, I, f_1, f_2\right) = f_1\left(l, I\right)^T f_2\left(l, I\right). \tag{2}$$

In the previous equation, f_1 and f_2 are the feature mapping functions, and the initial features I are mapped to $F1, F2 \in R^{C\times N}$, respectively. In this way, bilinear models are used to make full use of the fine-grain information under different receptive fields. We define the bilinear fusion based on the two branch features as

$$B = F1 \cdot F2^T, \tag{3}$$

where $B \in R^{C\times C}$ and $B_{ij} = \sum_{k=1}^{N} F1_{ik} \cdot F2_{jk}$. The bilinear image feature B integrates the details of person images between different channels, so we generate a set of channel attention $M \in R^{C\times C}$ through the Softmax layer after the bilinear model:

$$M_{i,j} = \frac{\exp\left(B_{ij}\right)}{\sum_{i=1}^{C} \exp\left(B_{ij}\right)}. \tag{4}$$

$M_{i,\ j}$ represents the influence of channel j on channel i. Therefore, attention M can be regarded as a summary of the interchannel dependency and the local feature strength of a given feature graph. In order to achieve the effect of attention mapping on the original features, we will pay attention to the transposition of the torque matrix M to the original features I_1. The results are reshaped into three-dimensional space with the same shape as input tensor to get the output features $D \in R^{C\times H\times W}$:

$$\begin{aligned} D_1 &= M^T \cdot I_1, \\ D_1 \in \mathrm{R}^{C\times N} &\xrightarrow{\text{reshape}} D \in \mathrm{R}^{C\times H\times W}. \end{aligned} \tag{5}$$

Finally, the pointwise convolution (PW 1×1) is used to fuse the image features of the channel attention information to form a complete bottleneck module.

2.2. Hard Sample Selection Mechanism. The selection of training samples by traditional recognition methods mostly put all the positive and negative samples of batch into training. Although the number of samples is large, it is difficult to avoid the fact that the sample is too simple, which is of little significance to model training and waste of computing resources. For example, through continuous person images, we could get similar or even identical sample images. However, due to the large difference of clothing, shape, or background in some person images, the negative samples are relatively simple. The model can easily distinguish person differences. Therefore, a large number of simple samples will waste a lot of time and computing resources in the later stage of the model.

To solve this problem, a simple hard sample mining module is designed to extract hard samples from all the samples into the initial P2G similarity optimization module [21], so as to enhance the ability of network learning complex samples, thereby improving the generalization ability of the model. As shown in Figure 4, the same colour features represent the same person ID, while different colour features represent different person IDs.

In order to optimize the network training stage, the ID of the image is applied as a known parameter to the training stage, in the hard sample mining module, to ensure that the network can effectively learn the accurate feature expression from the hard sample image pair. In the training stage, each batch has n different person IDs, and each person ID has K different pictures; that is, each batch contains $n \times K$ images. Suppose that every ID selects a picture as the probe; then the whole group of images is input into the improved Siamese Network module, and the image features are extracted and the similarities between each probe and other images (Gallery) are calculated, so as to get the initial P2G similarity score vector. According to the initial similarity score, we can select the difficult samples in the whole group of pictures for each probe picture. We take the samples with lower similarity scores under the same ID as hard positive samples and the samples with higher scores under different IDs as the difficult negative samples. Next, the hard positive samples and the hard negative samples are used as the input of the P2G similarity update module.

We set the hard sample mining module before the P2G similarity optimization module, which is equivalent to a mask for the initial similarity matrix S, making the similarity information of the simple samples inactive automatically, and the hard samples as the input of the similarity update module. In this way, the similarity information of hard samples is used to optimize the initial P2G similarity, so as to fully exploit the correlation between difficult samples under limited computing resources.

2.3. P2G Similarity Optimization and Feature Fusion. Due to the fact that the local part of person image contains rich person details, common person reidentification methods based on local images (such as key points detection, image segmentation, person attributes, etc.) require expensive semantic component learning process. In the form of feature partitioning, we design a local and global feature similarity fusion module to make full use of the detailed information of the image features, while avoiding the cumbersome process of local feature extraction, which also realizes the automatic perception of the characteristics of effective parts [22].

We adjust the input image to a size of 256×128, so after the ResNet50 network, the output of the last layer is $16 \times 8 \times 2048$. Feature grouping module is set at the last layer of the network output. The output feature is grouped in the

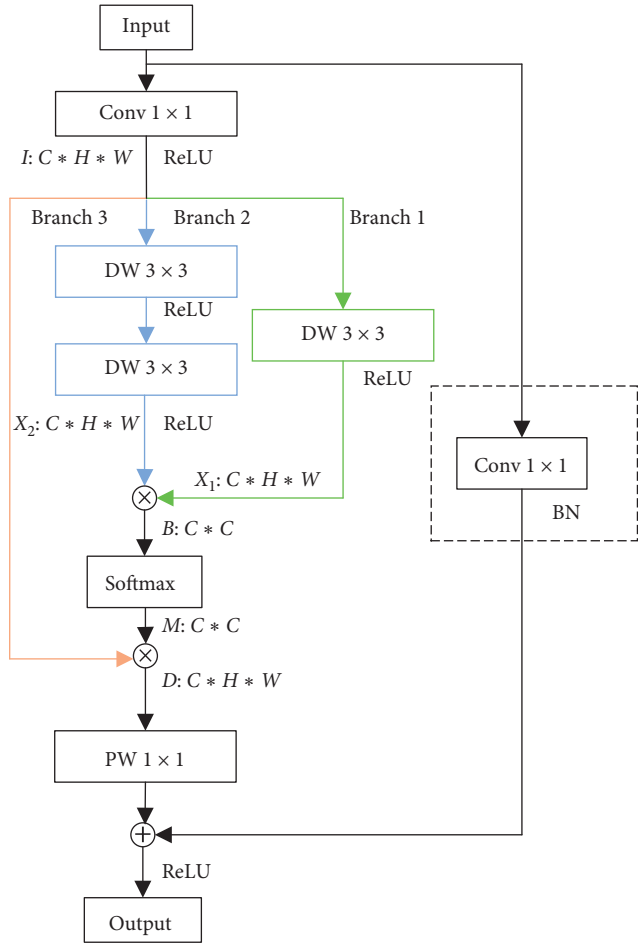

Figure 3: Bilinear channel fusion attention module.

way of horizontal overlap into the 3 upper, middle, and lower groups of size of $8 \times 8 \times 2048$. The principle of feature grouping is shown in Figure 5.

After grouping, each feature is calculated separately by local feature distance, and the local similarity score is obtained at the same time, and the local similarity is updated by the similarity optimization module [21]. Adding to the global similarity estimation, we get 4 sets of similarity scores. However, the 3 sets of local similarities make full use of local detail feature information, expressing the local correlation between images and showing the similarity is based on fine-grained image and locality. The global similarity grasps the overall style and characteristics of the image. It has a complete expression to the similarity relation of the person's overall contour and attributes. Each group of similarity has its own advantages, but it cannot fully express the correlation between images. After obtaining the similarity measure results, we integrate the local and global similarity to get the complete similarity measure results. The similarity module and the global similarity module are shown in Figure 6.

In the similarity fusion module, the 4 groups of P2G similarity scores are first spliced into a set of similarity score vectors through flatten and concat. Then they are input to the fully connected layer. The FC layer takes all the similarity scores extracted from the previous layer as input features and maps them to a new dimension vector as the output value in the fully connected mode. The result of similarity measurement after fusion is obtained.

The FC layer cannot only map the learned distributed feature representation to the corresponding space in the network but also realize the learning of distributed feature weights through the training process. In forward propagation, a fusion map of similarity scores is achieved through a fully connected layer. When being back-propagated, the weights of different local feature similarity and global feature similarity are automatically adjusted through training, so as to realize automatic perception of local details and global information of the image. After multiple iterations of network training, the P2G similarity is gradually refined, and the accuracy of pedestrian reidentification ranking results has been correspondingly improved.

3. Experiment and Test

3.1. P2G Similarity Update Algorithm. In this paper, Market-1501 and CUHK03 datasets are used to carry out experimental and comparative analysis of two datasets with larger data volume.

The Market-1501 dataset was collected in real time on the campus of Tsinghua University in the summer of 2015. The

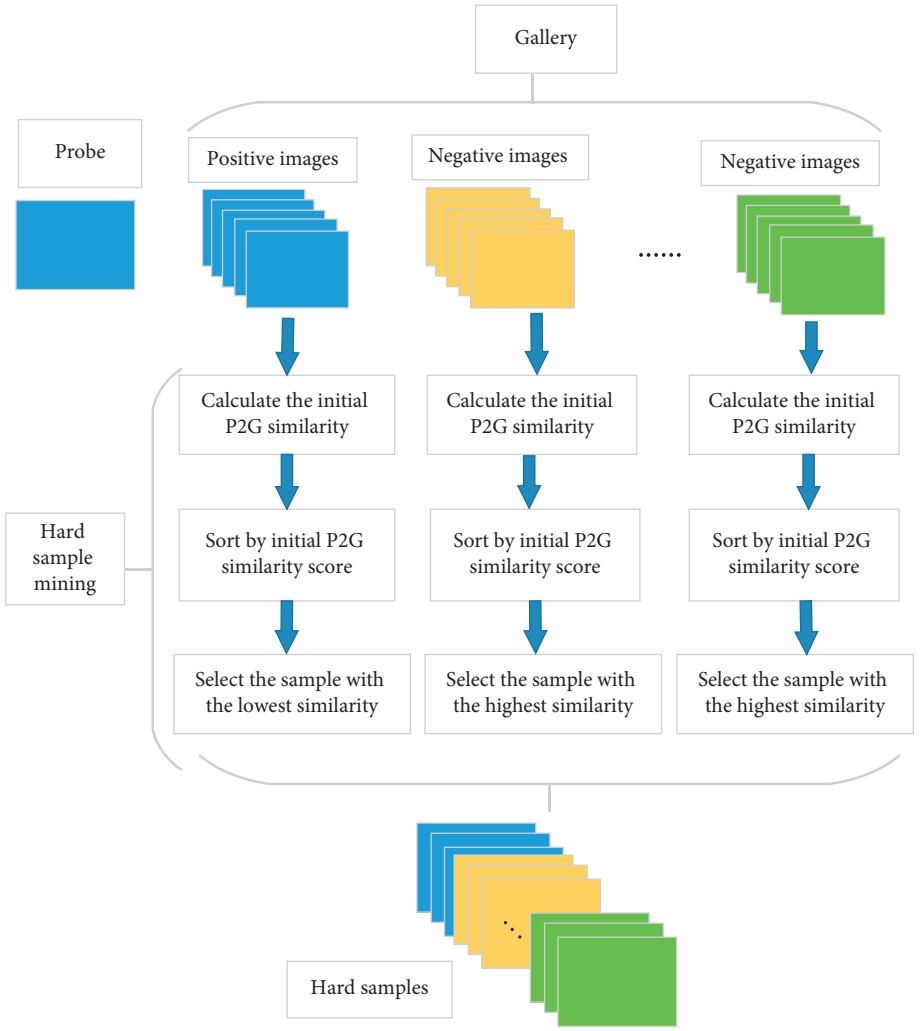

Figure 4: The diagram of hard sample mining module.

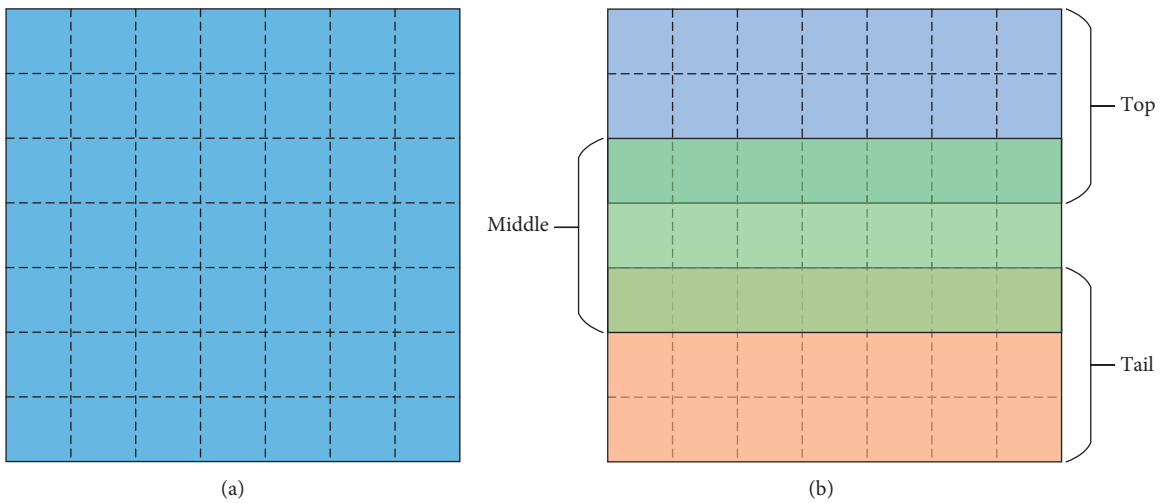

Figure 5: (a) Original feature; (b) features grouped in the way of horizontal overlap.

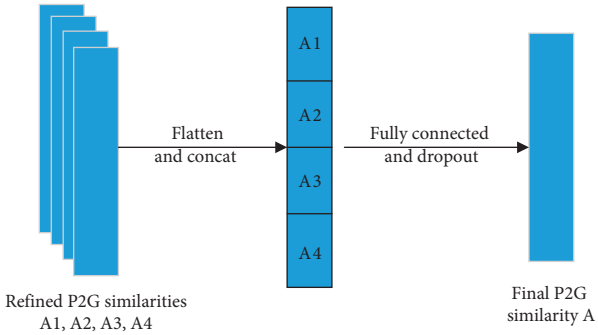

Figure 6: Local and global feature similarity fusion module.

person images came from 6 different cameras, including 32668 pictures containing 1501 ID persons. We divide the 1501 persons into two groups: training set and test set: the training set contains 751 people and 12936 pictures, and the test set contains 19732 pictures of 750 people.

The CUHK03 dataset was collected on the campus of Chinese University of Hong Kong. The images came from 10 different cameras. There are 14097 images of 1467 persons, with an average of 9.6 training pictures per person. The dataset is divided into two groups, of which 767 are training sets and 700 are test sets. At the same time, the dataset also provides two-person-marking information detected by manual annotation and DPM detector.

In this paper, we use the cumulative matching characteristic (CMC) curve and the mean Average Precision (mAP) to evaluate the performance of two models.

Among them, the cumulative matching characteristic (CMC) curve takes Rank-n as the abscissa and Rank-n matching result accuracy rate as the ordinate. Rank-n represents, according to P2G similarity score, the ratio of the number of the correct matching probes to the number of all the probes, that is, the front n hit rate. In the CMC curve, we often use three sets of results of Rank-1, Rank-5, and Rank-10 to visualize the effect of the model.

In this paper, we propose a person reidentification model experiment using RTX 2070 graphics acceleration calculation. Under the Anaconda development environment, we implement it based on PyTorch open source framework and Python programming. Before model training, we pretrained the ResNet50 parameters on the ImageNet dataset. In order to improve the expressive power and generalization ability of the model, we set up the data preprocess before training the model. We used random erasure, random horizontal flipping, and data standardization to preprocess the image. The size of the image was adjusted to 256×128, with the mean [0.485, 0.456, 0.406] and variance [0.229, 0.224, 0.225] normalizing the RGB three-channel image as the input of the network.

There are two key elements in the person reidentification model: (1) Improve the ResNet50 based Siamese Network to extract robust person features, so as to improve the efficiency of similarity measurement, to achieve the purpose of improving the accuracy of reidentification. (2) Use the hard sample mining and similarity optimization fusion module to refine the initial similarity results and achieve the optimization of similarity ranking results. In the experiment, we not only integrate two key elements into the whole model training to verify the overall effect of the model but also conduct ablation experiments on two modules to observe the effectiveness of each independent part.

3.2. Improving the Siamese Network Module Experiment and Result Analysis. In order to verify the effectiveness of the improvement in Siamese Network, we extract the person features based on the improved network structure. Then Euclidean distance is used to measure the similarity between the probe and the gallery, and the image in the gallery is sorted according to the similarity scores. In addition to the contrast experiment, based on the original ResNet50 and the improved ResNet50, we also carried out ablation experiments on two kinds of training methods, namely, the contrastive loss based on the Euclidean distance and the classification loss based on cross-entropy. The experimental results of the improved Siamese Network module on Market-1501 are shown in Table 1.

From Table 1, we can see that, in the improved Siamese Network module using Euclidean distance metric, the average accuracy of the model is increased by 1.7 percentage points. Rank-1 is also increased by about 12.74%. After adding the cross-entropy loss, the improved Siamese Network is a backbone network and the first hit rate of Rank-1 is 89.86%. The average accuracy index is also improved to 73.64%. The improved Siamese Network module is used to extract the features. The results are all higher than the Siamese Network model based on the original ResNet50. This indicates that the improved ResNet50 can extract more robust person features when using the simplest similarity measure and classification model and improve the accuracy of person reidentification.

Based on the calculation results of Rank-n, we draw the CMC curve of the person reidentification of two backbones under different training methods, as shown in Figure 7. In the figure, the CMC curve of the improved Siamese Network model is always located above the CMC curve of the original ResNet50, which indicates that the hit rate of person ranking results from the improved model of this paper is always higher than that of the original model.

3.3. P2G Similarity Optimization Experiment and Result Analysis. We not only did ablation experiments on every part of the module but also compared the results with other algorithm models to see the effectiveness of P2G similarity optimization module guided by G2G similarity.

"Baseline" in Table 2 refers to the backbone based on the original ResNet50 network. The Euclidean distance is used to measure the initial similarity, and the cross-entropy is used to carry out the classification loss of the network. From Table 2, it can be found that, after the hard sample mining module is added to the model, the average precision mAP increased to 79.15%. After the local feature division and the integration of global and local feature similarity, the mAP index of the model reached 82.5%, and the Rank-1 rate increased to 92.3%.

Table 1: Comparison of experimental results before and after the Siamese network improvement.

Network settings	mAP	Rank-1	Rank-5	Rank-10
ResNet50 + Euclidean	40.54	60.12	78.02	85.34
Improved ResNet50 + Euclidean	42.30	72.86	87.41	90.68
ResNet50 + Euclidean + cross-entropy	71.59	88.74	95.22	96.97
Improved ResNet50 + Euclidean + cross-entropy	73.64	89.86	96.74	98.21

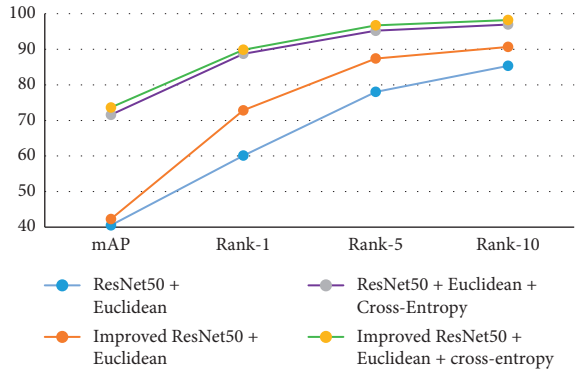

Figure 7: The CMC curves of different trunk network models.

Table 2: Comparison of ablation experimental results of similarity optimization module.

Network setting	mAP	Rank-1	Rank-5	Rank-10
Baseline + P2G optimization	78.28	89.97	96.35	97.62
Baseline + P2G optimization + hard sample mining	79.15	90.73	96.92	98.04
Baseline + P2G optimization + hard sample mining + feature fusion	82.5	92.3	97.1	98.3

In Figure 8, we draw the CMC curve of the model under different conditions. After introducing the hard sample selection mechanism and the global and local similarity design, the hit rate of the model is further improved. The CMC curve of the improved model is obviously higher than that of the basic mode curve.

3.4. Overall Model Training and Experimental Result Analysis. In addition, we have integrated the design of the model and took the improved ResNet50 as the backbone. Based on the P2G optimization model, the hard sample mining mechanism is introduced, and feature groups and similarity fusion modules are implemented to achieve the accuracy of person image similarity. The performance of the overall model is compared with the existing models on the Market-1501 dataset and the CUHK03 dataset. The results are shown in Tables 3 and 4.

From the experimental results on the Market-1501 dataset, we can see that, compared with the LSTM network that integrates the Long Short-Term Memory into the Siamese Network and the TriHard algorithm, this model has a great advantage in mAP and Rank-1 accuracy. It is also in common with the model Spindle based on human image segmentation and component alignment. Compared with Spindle Net, AlignedReID, and GLAD models, this method has a significant improvement in average precision and Rank-n rate. Compared with the current popular reranking methods, K-reciprocal and PSE + ECN, although the average

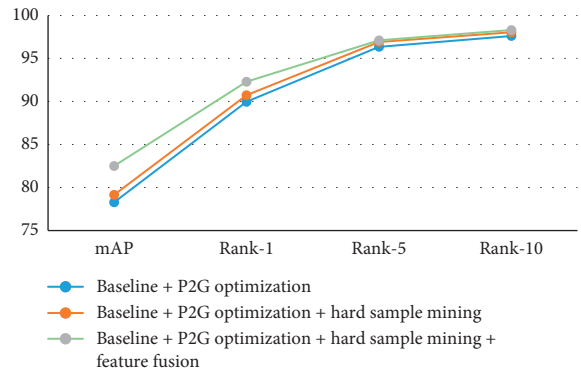

Figure 8: The CMC curves of similarity optimization module.

Table 3: Experimental results of different models on the Market-1501 dataset.

Method	mAP	Rank-1	Rank-5	Rank-10
LSTM	35.3	61.6	—	—
Spindle Net	—	76.9	91.5	94.6
TriHard	69.1	84.9	94.2	—
GLAD	73.9	89.9	—	—
K-reciprocal	63.6	77.1	—	—
AlignedReID	79.3	91.8	—	—
PSE + ECN	84.0	90.3	—	—
Ours	82.7	92.5	97.8	98.6

"—" indicates that the index data are not given in the original paper of the method.

Table 4: Experimental results of different models on the CUHK03 dataset.

Method	mAP	Rank-1	Rank-5	Rank-10
LSTM	—	57.3	80.1	88.3
Spindle Net	—	88.5	97.8	98.6
Quadruplet	—	75.5	95.2	99.2
GLAD	—	82.2	95.8	97.6
K-reciprocal	67.6	61.6	—	—
AlignedReID	—	92.4	98.9	99.5
SVDNet	84.8	81.8	95.2	97.2
Ours	92.0	94.3	98.7	99.3

"—" indicates that the index data are not given in the original paper of the method.

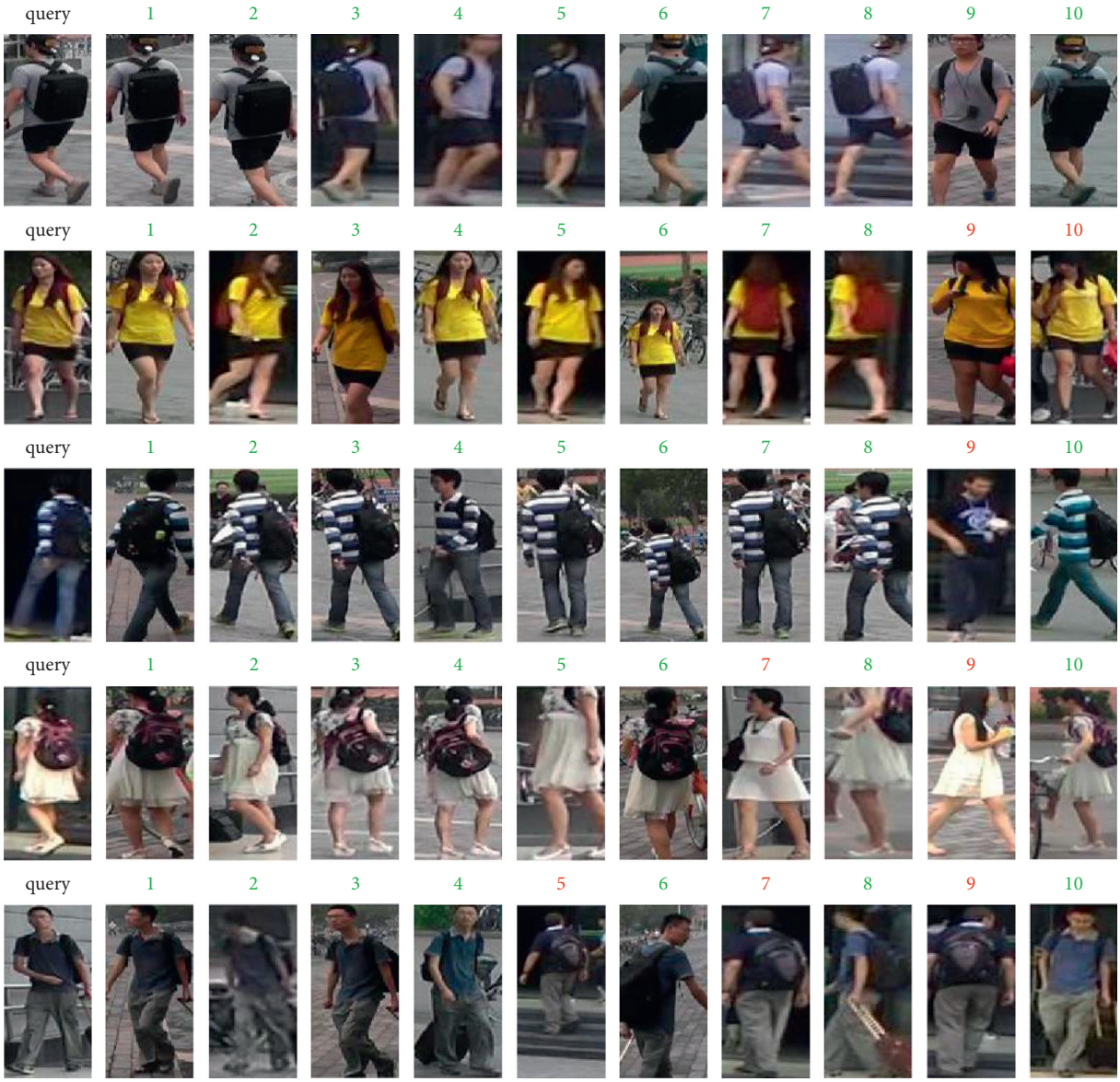

Figure 9: The result of the overall model.

precision index is slightly lower than the PSE + ECN method, the Rank-1 rate is significantly higher than the two.

From the experimental results in Table 4, we can find that, compared with the current popular algorithm models, the average precision mAP of this model on the CUHK03 dataset is excellent, and the accuracy of Rank-*n* is also good. Combined with the experimental results of Table 3, it is found that this method achieves high accuracy on two

common large-scale datasets of Market-1501 and CUHK03. It is shown that the model proposed in this paper has stronger recognition ability in person reidentification tasks.

Figure 9 is a person reidentification test result on the Market-1501 dataset. The first column is the query image, and the right side is the result of re-recognition for persons. According to the similarity score with the query image, we sequentially sort the retrieval results from high to low. The sorting of correct query result, that is, the same ID with query image, is marked in green and in red otherwise.

4. Conclusions

In view of the large difference of person under different visual fields, in order to fully extract effective person details information and to learn robust person expression and solve the key problem of reidentification task, this paper proposes a person reidentification algorithm, which includes three main aspects. Firstly, a bilinear channel fusion attention mechanism is designed to improve the bottleneck of ResNet50, which realizes the convergence of image features on different channels under the multireceptive fields and enhances the learning of image feature details in the form of attention, so as to improve the ability of expressing fine-grained information in pedestrian images. Secondly, the hard sample mining mechanism is designed, and the hard sample is used as the object of similarity optimization to reduce the initial P2G optimization model parameters. At the same time, it can reduce the computing resources and enhance the generalization ability of the model. Finally, the similarity optimization module is introduced to realize the automatic perception of the key parts of the image by the fusion of grouping and global similarity, so as to achieve more accurate and efficient person reidentification. The paper also carries out a reidentification experiment for the proposed algorithm, compares it with the mainstream algorithm, and obtains a good result. The experiment shows that the algorithm in this paper has a relatively obvious improvement in the average accuracy and Rank-n and has an obvious advantage in Rank-1. The next step will be to comprehensively improve performance indicators and network light.

Data Availability

The image datasets used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the following projects: Scientific Research Project of National Language Commission (YB135-125), Key Research and Development Project of Shandong Province (2019GGX101008 and 2016GGX105013), Natural Science Foundation of Shandong Province (ZR2017MF048), and Science and Technology Plan for Colleges and Universities of Shandong Province (J17KA214).

References

[1] D. Gray, S. Brennan, and H. Tao, "Evaluating appearance models for recognition, reacquisition, and tracking," in *Proceeding of the 10th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS 2007)*, pp. 356–368, Rio de Janeiro, Brazil, October 2007.

[2] M. Farenzena, L. Bazzani, A. Perina, V. Murino, and M. Cristani, "Person re-identification by symmetry-driven accumulation of local features," in *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2360–2367, San Francisco, CA, USA, June 2010.

[3] S. Liao, Y. Hu, X. Zhu, and S. Z. Li, "Person re-identification by local maximal occurrence representation and metric learning," in *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2197–2206, Boston, MA, USA, June 2015.

[4] S. Wu, Y.-C. Chen, X. Li, A.-C. Wu, J.-J. You, and W. S. Zheng, "An enhanced deep feature representation for person re-identification," in *Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision*, pp. 743–758, Lake Placid, NY, USA, March 2016.

[5] Y.-L. Wei and C. H. Lin, "Efficient weighted histogram features for single-shot person re-identification," *Journal of Signal Processing Systems*, vol. 90, no. 4, pp. 477–491, 2018.

[6] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: person retrieval with refined part pooling (and A strong convolutional baseline)," *Computer Vision-ECCV 2018*, Computer Vision-ECCV 2018-15th European Conference, Munich, Germany, pp. 501–518, 2018.

[7] M. Kostinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2288–2295, Providence, RI, USA, June 2012.

[8] S. Wang, T. Liang, Q. Chen, J. Na, and X. Ren, "USDE-based sliding mode control for servo mechanisms with unknown system dynamics," *IEEE/ASME Transactions on Mechatronics*.vol. 25, no. 2, pp. 1056–1066, 2020.

[9] Y. Wang, Z. Chen, F. Wu, and G. Wang, "Person re-identification with cascaded pairwise convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1470–1478, Salt Lake City, UT, USA, June 2018.

[10] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, "Person re-identification by multi-channel parts-based CNN with improved triplet loss function," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1335–1344, Las Vegas, NV, USA, June 2016.

[11] E. Ristani and C. Tomasi, "Features for multi-target multi-camera tracking and re-identification," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6036–6046, Salt Lake City, UT, USA, June 2018.

[12] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3492–3506, 2017.

[13] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1213–1225, Honolulu, HI, USA, November 2017.

[14] J. Na, Y. Huang, X. Wu, S.-F. Su, and G. Li, "Adaptive finite-time fuzzy control of nonlinear active suspension systems

with input delay," *IEEE Transactions on Cybernetics*, vol. 50, no. 6, pp. 2639–2650, 2020.

[15] W. Chen, X. Chen, J. Zhang, and K. Huang, "Beyond triplet loss: a deep quadruplet network for person re-identification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1320–1329, Honolulu, HI, USA, July 2017.

[16] S. Bai, X. Bai, and Q. Tian, "Scalable person re-identification on supervised smoothed manifold," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 674–689, Honolulu, HI, USA, July 2017.

[17] D. Li, X. Chen, Z. Zhang, and K. Huang, "Learning deep context-aware features over body and latent parts for person re-identification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 384–393, Honolulu, HI, USA, July 2017.

[18] S. Zhou, J. Wang, J. Wang, Y. Gong, and N. Zheng, "Point to set similarity based deep feature learning for person re-identification," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 452–465, Honolulu, HI, USA, July 2017.

[19] M. S. Sarfraz, A. Schumann, A. Eberle, and R. Stiefelhagen, "A pose-sensitive embedding for person re-identification with expanded cross neighbourhood re-ranking," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 420–429, Salt Lake City, UT, USA, June 2018.

[20] S. Bai, P. Tang, P. H. S. Torr, and J. Latecki, "Re-ranking via metric fusion for object retrieval and person re-identification," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 740–749, Long Beach, CA, USA, June 2019.

[21] G. Wang, S. Gao, and D. Fan, "A G2G similarity guided person re-identification algorithm," *Journal of Physics: Conference Series*, vol. 1453, no. 1, Article ID 012035, 2019.

[22] Y. Sun, Q. Xu, Y. Li et al., "Perceive where to focus: learning visibility-aware part-level features for partial person re-identification," in *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 820–834, Long Beach, CA, USA, June 2019.