Hindawi
Complexity
Volume 2021, Article ID 6673461, 10 pages
https://doi.org/10.1155/2021/6673461

Research Article

Person Reidentification Model Based on Multiattention Modules and Multiscale Residuals

Yongyi Li[1], **Shiqi Wang**[1], **Shuang Dong**[1], **Xueling Lv**[2], **Changzhi Lv**[1], **and Di Fan**[1]

[1]*Shandong University of Science and Technology, Qingdao, Shandong 266590, China*
[2]*Department of Management and Economics, Tianjin University, Tianjin 300072, China*

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

Received 25 November 2020; Revised 27 January 2021; Accepted 27 February 2021; Published 19 March 2021

Academic Editor: Rui Wang

At present, person reidentification based on attention mechanism has attracted many scholars' interests. Although attention module can improve the representation ability and reidentification accuracy of Re-ID model to a certain extent, it depends on the coupling of attention module and original network. In this paper, a person reidentification model that combines multiple attentions and multiscale residuals is proposed. The model introduces combined attention fusion module and multiscale residual fusion module in the backbone network ResNet 50 to enhance the feature flow between residual blocks and better fuse multiscale features. Furthermore, a global branch and a local branch are designed and applied to enhance the channel aggregation and position perception ability of the network by utilizing the dual ensemble attention module, as along as the fine-grained feature expression is obtained by using multiproportion block and reorganization. Thus, the global and local features are enhanced. The experimental results on Market-1501 dataset and DukeMTMC-reID dataset show that the indexes of the presented model, especially Rank-1 accuracy, reach 96.20% and 89.59%, respectively, which can be considered as a progress in Re-ID.

1. Introduction

As an important video intelligent analysis technology, person reidentification (Re-ID) uses computer vision technology to realize the identification and matching of target pedestrians in a multicamera network with non-overlapping fields of view. That is, given a pedestrian test image (Probe), the pedestrian image is retrieved under cross-monitoring equipment from all gallery images (Gallery) [1]. Compared with the perspective scenes of fixed monitoring equipment, person reidentification technology solves the problem of its visual limitations and can be well matched with pedestrian detection and pedestrian tracking scenes. It is a key technology for target tracking, urban intelligent security, and prevention and control of public places such as supermarkets, airports, stations, exhibition halls, and exhibition centers.

In recent years, research in the field of person rerecognition has focused on representation learning [2], metric learning [3], local features [4], video sequences [5], and Generative Adversarial Networks (GAN) generated images [6]. The performance of the person reidentification system is getting better and better, and it has made great progress. However, in the real scene, the problems of different image scales, low resolution, occlusion, and illumination differences that seriously affect the recognition effect are still not well resolved. Therefore, in recent years, many scholars have begun to study the attention mechanism and local features in order to improve the performance of the recognition model. At present, there has been some progress in the basic network and local features.

In terms of basic networks, most of models used ResNet50 as the backbone network, but also improved lightweight convolution models as the backbone network were also selected, such as Omni-Scale Feature Learning Network (OSnet) [7] and Robust-Re-ID [8]. The improvement ideas of these networks to the basic network mainly focus on the research and application of the attention mechanism. For example, Hou et al. proposed Interaction-and-Aggregation Network (IA-Net) [9], which constructs an

IA block by combining channel attention and spatial attention and then embeds it in the residual network to aggregate channel and spatial information. Chen et al. proposed the Mixed High-Order Attention Network (MHN) [10], which uses the high-order attention distribution High-Order Attention (HOA) module to obtain more feature information. Chen and Ding et al. proposed Attentive But Diverse Network (ABD-net) [11], which combines channel attention and position attention in parallel and adds them to the local feature network for feature fusion. Xia et al. proposed Second-Order Nonlocal Attention (SONA) network [12], which adds covariance matrix to the first-order nonlocal attention structure to make it second-order structure. This method effectively enhances the information flow between residual convolution blocks.

In terms of local features, image horizontal block is a common step in local feature extraction [4]. However, the disadvantage is that the requirements for image alignment are relatively high. If the two images are not aligned up and down, it is likely that body parts will be misplaced and compared, such as head and background contrast, which will increase the probability of model judgment error. Therefore, Zhang et al. designed a dynamic alignment network AlignedReID [13], which can automatically align image blocks from top to bottom without additional information. Some literatures used some prior knowledge to align pedestrians. For example, Zhao et al.'s spindle net [14] first estimated the key points of travelers with the attitude estimation model and then used affine transformation to align the same key points. In addition, Sun et al. proposed a Part-based Convolutional Baseline (PCB) [15] method to divide the feature map into six blocks horizontally and used refined part pooling (RPP) method for local alignment. Later, some researchers found that the combination of global features and local features can improve the expression ability of the network, and the local features are divided more carefully. Wang et al. proposed Multiple Granularities Network (MGN) [16] based on discriminative features, which uses a more detailed combination of local features and global features and achieves quite good recognition results. In addition, Zheng et al. [17] proposed the pyramid block model, which integrates the local and global information and the progressive clues between them and solves the occlusion problem to a certain extent.

However, there are still some problems to be solved in the above-mentioned attention mechanism and local features research. For example, some attention mechanisms need multiple matrix calculations, and the coupling with the original network is not ideal when joining the basic network. In addition, in the segmentation of feature map, the more blocks, the better local feature expression, but it will increase the amount of model parameters. If there are few blocks, the recognition rate will not be improved. In this regard, we designed a person reidentification model based on attention fusion and multiscale residuals. The model mainly solves the problems of poor coupling between attention mechanism and original network and scientific expression of local features. The model uses an improved ResNet50 as the backbone network and is designed with global and local branch structures. The paper added Combined Attention Fusion Module (CAFM) and Multiscale Residual Fusion Module (MSFM) to the original ResNet50 [18] to effectively concatenate the feature information between residual blocks and better integrate multiscale features. The global branch uses a Dual Ensemble Attention Module (DEAM) to enhance the network's channel aggregation and location awareness capabilities. The local branch is divided into fine-grained features by multiproportion block method to further refine the local features. In the experiment, the network in this paper has achieved good results on the Market-1501 and DukeMTMC-reID datasets, and the indicators are better than other Re-ID networks.

2. Person Reidentification Model Based on Attention Fusion and Multiscale Residuals

The algorithm model framework of this paper is shown in Figure 1. Firstly, the improved ResNet50 network and local and global branches are used to extract pedestrian features of Probe and Gallery, respectively. Then the similarity between Probe and Gallery pedestrian features is calculated. Finally, the similarity scores are sorted to obtain the retrieval results of all the images of Probe in the Gallery. The model in this paper is based on the ResNet50 network with multiple improvements and model extensions to enhance feature extraction and expression capabilities and effectively improve the recognition rate.

The feature extraction network designed in this paper is shown in Figure 2. Its backbone network is an improved ResNet 50, and it is equipped with global attention branch and local fine-grained feature branch. For the improvement of the backbone network, Combined Attention Fusion Module (CAFM) is added after ResNet50 Stage1 to strengthen the flow of feature information between residual blocks. Multiscale Residual Fusion Module (MSFM) is added after Stage4, which can perform multiscale feature extraction and selective fusion of the original residual features. For the design of the branch structure, the Dual Ensemble Attention Module (DEAM) is introduced into the global branch to strengthen the fusion of channel attention and position attention. Multiproportion block method is used to optimize the expression of fine-grained features in local branches.

2.1. The Improved ResNet 50 Network. ResNet 50 contains a total of 50 convolutional layers, which are input layer, output layer, and 48 hidden convolutional layers. 48 hidden layers are divided into four stages in the form of $3+4+6+3$ convolution residual bottleneck. In this paper, CAFM and MSFM are introduced into the backbone network for improvement, and the step size of the downsampling convolutional layer in Stage4 is changed from 2 to 1.The following is a detailed description of CAFM and MSFM.

2.1.1. Combined Attention Fusion Module (CAFM). The attention mechanism is a very important and effective method in deep learning [19]. Its essence is to linearly weigh

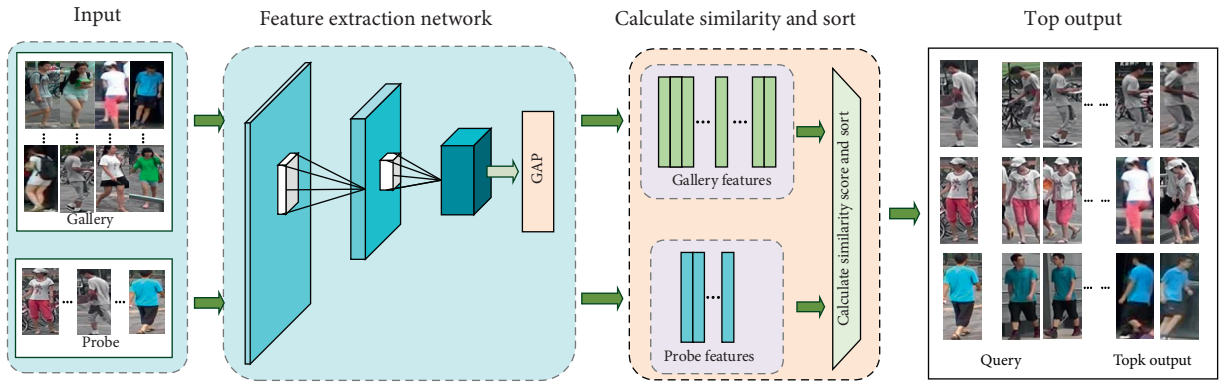

Figure 1: The overall framework of the algorithm model.

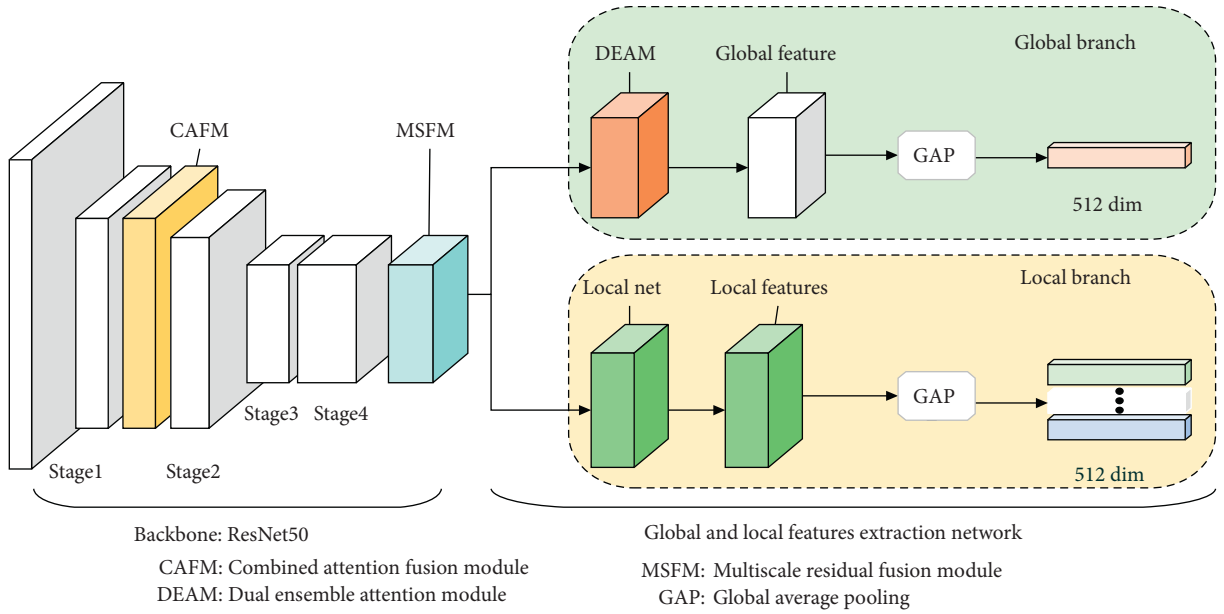

Figure 2: Schematic diagram of feature extraction network structure.

the relationship between things to obtain a new representation. In the field of person reidentification, the attention mechanism is often used to focus the attention of the network on the pedestrian's body, thereby eliminating the influence of factors such as background and occlusion. We designed the combined structure of Channel Attention Module (CAM) [11] and Nonlocal Attention Module (NAM) [20] and embedded it in the residual structure, which can fully concatenate the information between the residual blocks and increase the network's attention to the target feature. The specific structure of the module is shown in Figure 3.

In Figure 3, the input characteristics firstly fuse the channel information through the channel attention module (CAM), and the important location is perceived by the nonlocal attention module (NAM), and finally the attention information features integrated with the original input features. The Channel Attention Module is shown in Figure 3(b). It integrates all the relevant features in the channel map and selectively strengthens the correlated channel map. Nonlocal Attention Module (NAM) is shown in Figure 3(c). In theory, NAM is a position attention mechanism. Each position value of its output is a weighted average of other position values, which represents the dependence between pixels and other pixels. Among them, the function of softmax is to map the feature value between 0 and 1 to get the attention map.

2.1.2. Multiscale Residual Fusion Module (MSFM). Multiscale Residual Fusion Module is a dynamic selection mechanism that enables each neuron to select different receptive fields according to the size of the target feature [21].The Multiscale Residual Fusion Module structure designed in this paper is shown in Figure 4(b), which is mainly divided into two parts: multiscale feature extraction and feature selective fusion. The multiscale feature extraction part is mainly used for convolution extraction with different sizes of convolution kernels [22]. Taking into account the requirements of parameter weights and network performance, the module of this paper selects three sizes of convolution kernels of 3×3, 5×5, and 7×7

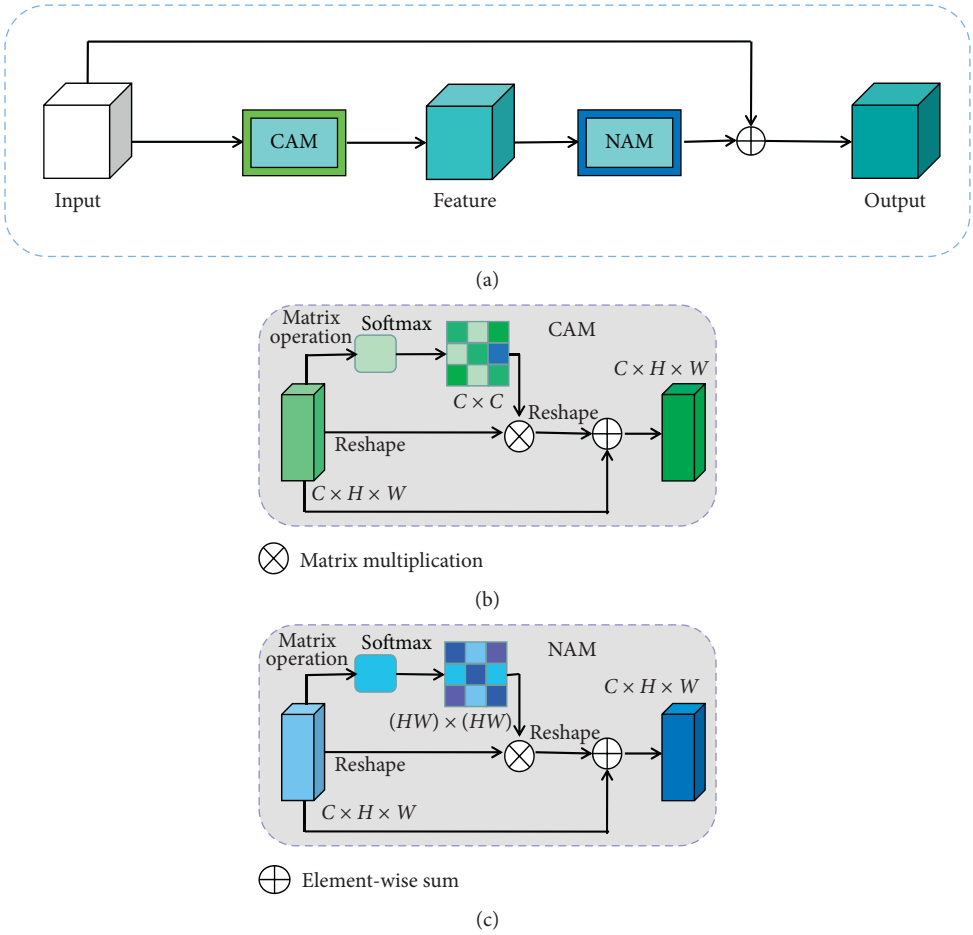

Figure 3: Structure of Combined Attention Fusion Module. (a) CAFM. (b) CAM. (c) NAM.

to extract features by group convolution. Feature selective fusion part is to fuse multichannel information for weight selection [20]. According to the selected weights, the feature images of convolution kernel with different sizes are fused.

The Multiscale Residual Fusion Module adopts a gate mechanism to control the information flow into the different branches of the next convolutional layer [23]. This mechanism fuses the information of all branches to realize the adaptive adjustment of the receptive fields of different sizes of neurons. This module first performs simple pixel-level addition and fusion of multibranch features to obtain feature $U \in R^{C\times H\times W}$, as shown in formula (1), where $U_r, U_f, U_s \in R^{CHW}$ are the output features of the three convolution channels.

$$U = U_t + U_f + U_s. \tag{1}$$

U uses global average pooling to encode global information to generate statistical information $S \in R^C$ on the channel. The c-th element S_c in S is obtained by compression calculation on the $H \times W$ dimension of U.

$$S_c = \frac{1}{H \times W} \sum_{i=1}^{H} U \sum_{j=1}^{W} (c, i, j). \tag{2}$$

Then normalize and nonlinearly operate on S to produce a compact feature $Z \in R^{C\times 1}$, which is obtained through a fully connected layer. Then softmax is operated on Z on the channel to get the soft attention information $\alpha, \beta, \gamma \in R^{C\times 1}$ between the three branch channels. Finally, α, β, γ are multiplied and fused together with U_t, U_f, U_s in the channel dimension to obtain a multiscale fusion feature. The calculation formula is as follows:

$$\begin{aligned} \alpha &= e^{AZ}/E, \\ \beta &= e^{BZ}/E, \\ \gamma &= e^{CZ}/E, \end{aligned} \tag{3}$$

$$U' = \alpha \cdot U_t + \beta \cdot U_f + \gamma \cdot U_s. \tag{4}$$

Among them, $E = e^{AZ} + e^{BZ} + e^{CZ}$; matrices $A, B, C \in R^{C\times C}$ represent the weight matrix of three branches, respectively, which are used to selectively fuse different scale features.

2.2. Global Feature Extraction Network Based on Dual Ensemble Attention Module. The combination of global features and local features is a common feature expression in pedestrian recognition network in recent years. The global branch of this paper uses the Dual Ensemble Attention

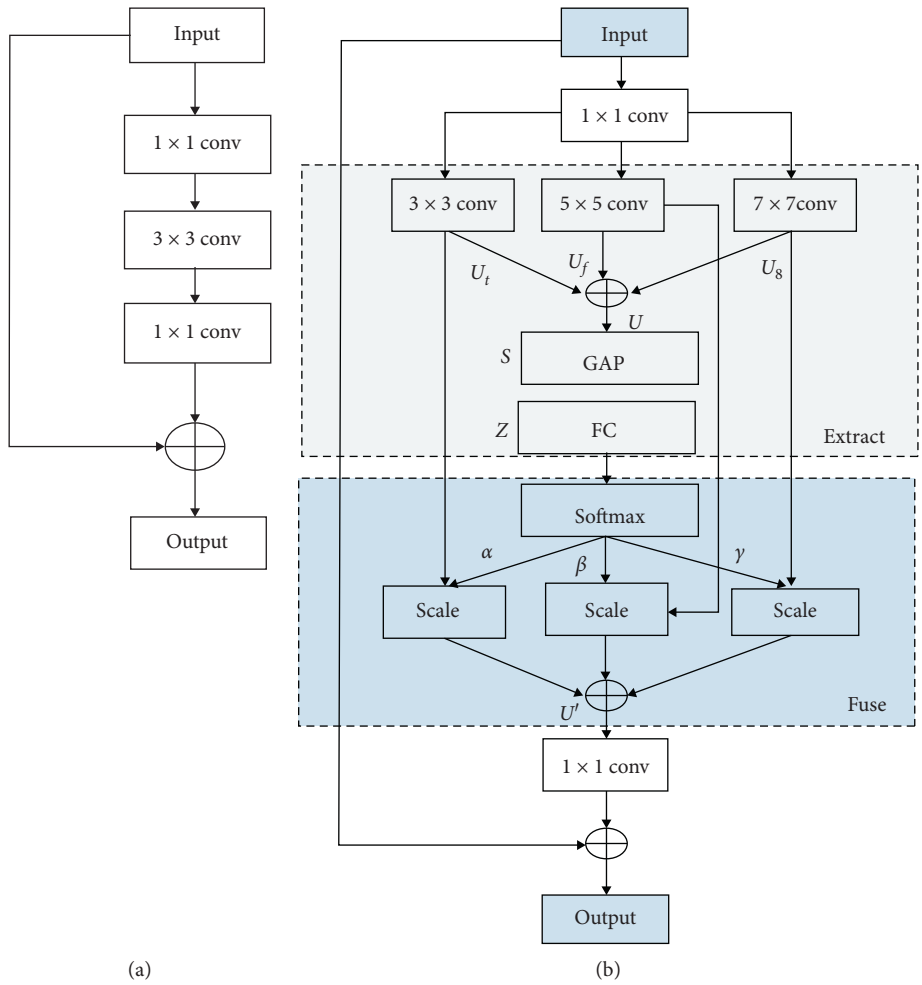

Figure 4: Structure diagram of Multiscale Residual Fusion Module. (a) Original residual structure. (b) Multiscale residual structure.

Module (DEAM) to further enhance the global features in space and channel dimensions on the basis of the output characteristics of the backbone network [24].

Based on the idea of ensemble learning [25], global branch designs two pairs of CAM and NAM into two-stage deep attention module [26]. As shown in Figure 5, this module models the semantic correlation information of spatial dimension and channel dimension, respectively. In the first stage, two basic CAM and NAM are integrated to realize the preliminary extraction of channel and position features. In the second stage, two improved CAM and NAM are integrated, and channel attention and position attention of the first stage are integrated, respectively. Through the weighted connection with the first stage, the second stage collects the information of attention block in the first stage, which further strengthens the learning ability of attention module. In the second stage, the information weighted fusion mode is shown in equation (5), where μ and τ are manually set parameters.

$$\begin{cases} M = \mu * M1 + \tau * M2, \\ N = \mu * N1 + \tau * N2. \end{cases} \tag{5}$$

2.3. Local Feature Based on Multiproportion Block and Reorganization. The design of local branch network is mainly to strengthen the expression of local features. This paper uses multiproportion block method to refine the fine-grained features and selects the part information with obvious features to enhance the network expression. What is different from the past is that we adopt a multiproportion block method to feature reorganization. In this method, the information of important parts is reused through block reorganization, and the information of secondary parts is weakened or discarded in varying degrees [27]. This design is derived from the observation of a large number of pedestrian pictures in reality and public datasets. It is found that the upper body features in the pictures are significantly stronger than the lower body. For example, the upper body has important fine-grained features such as human faces, hair, clothes logos, and hats, but the lower body has only monotonous legs and shoes, and there is no obvious distortion with the change of posture. In terms of occlusion, the lower body is the easiest to be occluded, such as people riding a bicycle, carrying a handbag to block their legs, and walking and being blocked by lawns and motor vehicles. If the

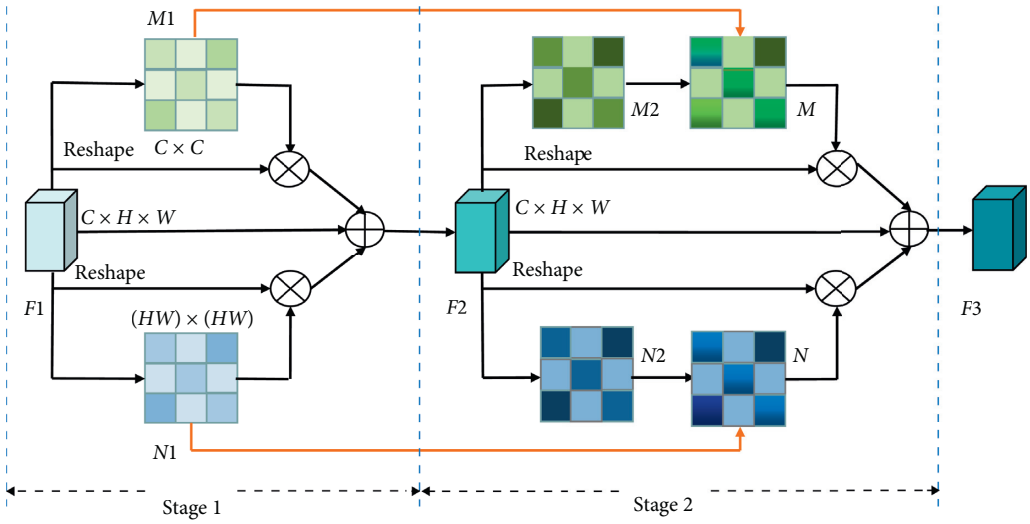

Figure 5: Structure diagram of Dual Ensemble Attention Module.

occluded parts or weak features receive more attention, the rerecognition effect will be reduced.

Some literatures have given more proportion of block patterns, but the more blocks, the heavier the network burden, and the improvement of identification index is not obvious. After experimental analysis, we designed a more optimized multiproportion block reorganization method, and the principle is shown in Figure 6. We first cut the H dimension of the feature map horizontally with different proportions of 1/2, 2/3, and 3/4 and then select the top 1/2, bottom 1/2, top 2/3, and top 3/4 of the feature map to be expressed as the local feature in cooperation with the global branch. This method not only reflects the importance of the top body characteristics but also covers the characteristics of the lower body. Through the upper 1/2, top 2/3, and top 3/4 blocks, characteristics such as head, torso, hands, and accessories are strengthened many several times. The bottom 1/2 block contains the characteristics of the lower body such as legs and shoes.

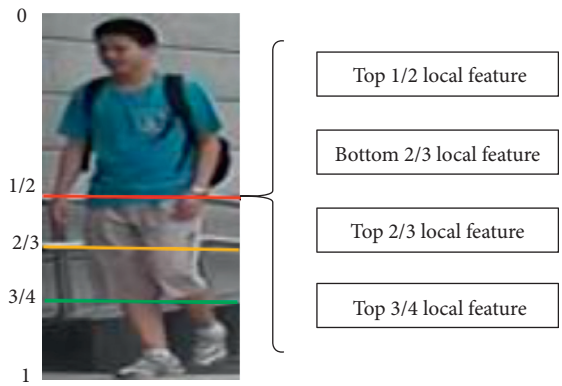

Figure 6: Schematic diagram of multiproportion block and reorganization.

3. Experiments

The person reidentification network model experiment proposed in this paper uses NVIDIA V100 16G graphics card to accelerate calculations in the CUDA10 environment and is implemented based on the PyTorch open-source framework and Python language programming.

3.1. Experimental Datasets and Their Extension. The experimental data in this article come from the Market-1501 dataset and DukeMTMC-reID dataset. The Market-1501 dataset was collected on the campus of Tsinghua University. The images come from 6 different cameras, one of which is of low resolution. The dataset contains a total of 32,668 pictures of pedestrians with 1501 IDs. The training set has 751 IDs and a total of 12936 images. The test set has 750 IDs and a total of 19,732 images. In all training sets, there are on average 17.2 pieces of training data for each ID. The DukeMTMC-reID dataset was collected at Duke University in the United States. The images came from 8 different cameras, and the borders of the pedestrian images were manually marked. The dataset provides training set and test set and has pedestrian attributes (gender, long and short sleeves, backpack, etc.) annotations. The training set contains 16522 images with a total of 702 IDs. On average, each ID has 23.5 training images, and the test set contains 17,661 images.

In order to improve the expression ability and generalization ability of the model, we expanded and preprocessed the dataset image, adopted random erasing, random horizontal flip, and data standardization [25], and adjusted the image size uniformly. The size is 384×128, and the RGB three-channel image is normalized with the mean value [0.485, 0.456, 0.406] and variance [0.229, 0.224, 0.225].

3.2. Experimental Parameters and Methods. In training, the ResNet50 model parameters pretrained on ImageNet dataset are used to initialize the model, and triplet hard loss with batch hard mining (TriHard loss) and cross entropy loss are used to accelerate convergence. The weight of triplet loss is

0.3, the margin is set to 1.2, and the weight of ID loss is 1, with label smoothing. In the model training, we first fix the initial weight of ResNet50 in the first 10 epochs and only train the weights of attention module and branch network and then continue training after releasing the last 90 epochs. The initial learning rate was 0.0003 when batch size was set to 32, and the learning rate decreased to 0.1 times at 30 and 60 epoch. In the experiment, Adam optimizer is selected to train the network. The momentum parameter is set to 0.9 and the weight attenuation coefficient is set to 0.0005. We also used label smoothing [28] and BNNeck [21] training methods.

In the test, the performance of the model was evaluated by two widely used indexes: Cumulative Match Characteristic (CMC) curve and mean average precision (mAP).

3.3. Experiment and Result Analysis. We use ResNet 50 as the baseline and then gradually add the designed modules for ablation experiments to observe the characteristics of the improvement measures of the model in this paper. The training and testing all adopt the aforementioned experimental parameters and methods. The added modules include Dual Ensemble Attention Module (DEAM), Combined Attention Fusion Module (CAFM), Multiscale Residual Fusion Module (MSFM), and local branch network (local-net). Among them, local-net (n-1)/n means that the local branch is divided into n blocks with different proportions, and the upper (n-1)/n part of the feature map is selected for local feature expression. We conducted experiments on the Market-1501 and DukeMTMC-reID datasets, respectively, and the comparison indicators were the mean average precision (mAP) and Rank-1 accuracy. The specific experimental results are shown in Table 1.

It can be seen from Table 1 that, on the Rank-1 accuracy of the Market-1501 dataset, the accuracy increased by 0.8% after adding DEAM based on the baseline (92.93%). On the basis of DEAM, we conducted more detailed experiments on local-net. It can be seen from the table that as the number of blocks increases, the accuracy rises. But when the number of blocks n is 5, the accuracy starts to decline, so our local-net selects the network structure when n is 4, and the Rank-1 accuracy reaches 95.57%. Then we added CAFM and MSFM to the comparative experiment based on the DEAM. It is found that the Rank-1 accuracy increases by about 0.5% after adding the two modules, but the mAP does not improve. After adding local-net, it is obvious that mAP and Rank-1 accuracies have increased significantly, which fully shows the importance of local branch. Finally, we combined all the improvement methods, and the mAP reaches 88.18% and Rank-1 accuracy reaches 96.20%. On the DukeMTMC-reID dataset, the Rank-1 accuracy increases with the addition of the above modules, reaching 89.59%.

From the overall ablation experiment, it is found that, in the Market-1501 dataset, the improvement method that has the greatest effect on the index improvement is the local-net design. This fully shows the effectiveness of multiproportion block scheme and the importance of local features. In terms of other modules, although the DEAM module is not as obvious as local-net, the accuracy is also greatly improved. For CAFM and MSFM modules, the improvement effect of single module may not be obvious. But, after using them together, especially with the local branch design, the experiment has achieved good results. To sum up, the modules we designed are effective.

In order to investigate the role of each module more comprehensively, we conducted a series of experiments on the Market-1501 dataset and obtained the CMC curve in the range of Ranks-1-40 of each module network, as shown in Figure 7. The abscissa of the CMC graph represents the number of hits, and the ordinate represents the hit probability of each rank. The five curves in Figure 7 are the performance of the ResNet50-based baseline after adding DEAM, local-net, CAFM, and MSFM. It can be seen from Table 1 and Figure 7 that the design modules and practices in this article have an improved identification index. The model in this paper incorporates the advantages of DEAM, local-net, CAFM, and MSFM, and the overall performance is greatly improved compared to the pure ResNet50. With the increase of each module, the performance also has a certain degree of improvement.

3.4. Comparison with Other Re-ID Models. We have compared the model with the person reidentification models based on Stripe, GAN, and Global Feature and Attention [7] in recent years. The experimental results are shown in Table 2. The scatter plot of several network indicators based on the Market-1501 dataset is shown in Figure 8.

As can be seen in Table 2, the Rank-1 (96.2%) accuracy of this model on the Market-1501 dataset is the same as Robust-ReID and is better than several other network models. mAP (88.2%) is better than other network models, except that it is slightly lower than Robust-Re-ID, SONA, and ABD-net. On the DukeMTMC-ReID dataset, the Rank-1 (89.6%) accuracy is higher than other network models except Robust-Re-ID, but the mAP (76.4%) is slightly lower than several models. This may be because the local part of the block combination has more features and there is greater redundancy. In the similarity comparison, although the first hit rate is high, the index on the n-th hit rate is not high, which leads to lower mAP. We will continue to pay attention and study this issue. In addition, we also performed a Re-ranking [36] experiment on the model in this paper, and the results are listed in the last row of Table 2. The results showed that mAP and Rank-1 accuracies reached 94.6% and 96.5% on the Market-1501 dataset, respectively, and mAP and Rank-1 accuracies reached 89.5% and 91.9% on the DukeMTMC-ReID dataset, respectively.

At the same time, the advantages of this model can be clearly seen in Figure 8. The overall index is better than the other 18 network models except Robust-Re-ID, especially in the Rank-1 accuracy.

3.5. Visualization Experiment and Results. We select four query sets on the Market-1501 dataset and perform Re-ID feature extraction [37] and similarity matching experiments [38] in the Gallery library. The matching results are shown in

Table 1: Ablation experiment results of each module.

Method	Market-1501		DukeMTMC-ReID	
	mAP	Rank-1	mAP	Rank-1
Baseline	84.14	92.93	72.75	85.95
Baseline + DEAM	85.28	93.73	75.21	87.52
Baseline + DEAM + CAFM	85.02	94.17	74.85	87.36
Baseline + DEAM + MSFM	84.87	94.29	75.46	87.43
Baseline + DEAM + local-net (1/2)	83.97	94.71	72.58	86.67
Baseline + DEAM + local-net (1/2 + 2/3)	84.76	94.95	73.75	86.40
Baseline + DEAM + local-net	87.54	95.57	77.10	87.75
Baseline + DEAM + local-net (+4/5)	87.13	95.15	74.56	87.84
Baseline + DEAM + local-net + CAFM	86.87	95.84	75.75	88.36
Baseline + DEAM + local-net + MSFM	87.44	95.90	76.12	88.30
Baseline + DEAM + local-net + CAFM + MSFM	**88.18**	**96.20**	**76.35**	**89.59**

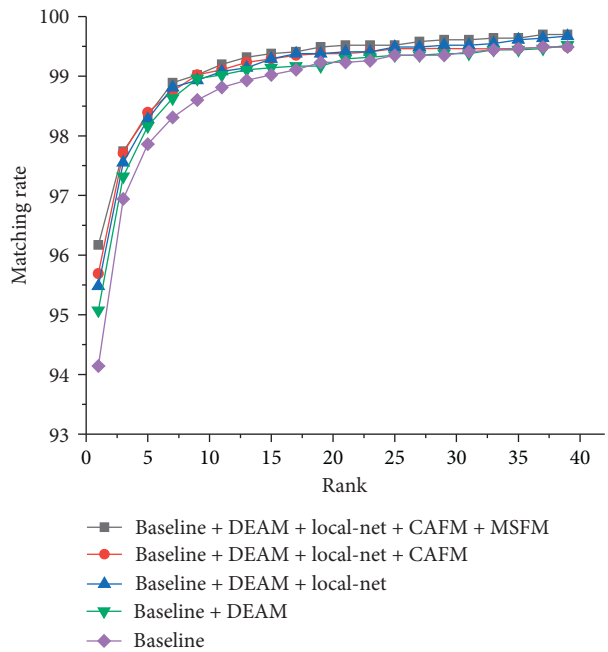

Figure 7: CMC curves of five networks on the Market-1501 dataset.

Table 2: Experimental results of this model compared with other Re-ID models.

Type	Method	Market-1501		DukeMTMC-reID	
		mAP	Rank-1	mAP	Rank-1
Stripe-based	AlignedReID [13]	77.7	90.6	67.4	81.2
	PCB + RPP [15]	81.6	93.8	69.2	83.3
	BFE [29]	85.0	94.5	75.8	88.7
	MGN [16]	86.9	95.7	78.4	88.7
	Pyramid [17]	88.2	95.7	79.0	89.0
	LocalCNN [30]	87.4	95.9	-	-
GAN-based	PN-GAN [31]	72.6	89.4	53.2	73.6
	DG-Net [32]	86.0	94.8	74.8	86.6
Global Feature	SVDNet [33]	62.1	82.3	56.8	76.7
	IA-Net [9]	83.1	94.4	73.4	87.1
	OS-Net [7]	84.9	94.8	73.5	88.6
	BagOfTricks [21]	85.9	94.5	76.4	86.4
	AGW [34]	87.8	95.1	79.6	89.0
Attention-based	MHN [10]	85.0	95.1	77.2	89.1
	ABD-Net [12]	88.3	95.6	78.6	89.0
	RGA-SC [35]	88.1	95.8	74.9	85.1
	SONA [12]	88.8	95.7	78.3	89.5
	Robust-ReID [8]	**89.7**	**96.2**	**80.3**	**89.8**
	Ours	88.2	**96.2**	76.4	89.6
	Ours(Re-rank)	**94.6**	**96.5**	**89.5**	**91.9**

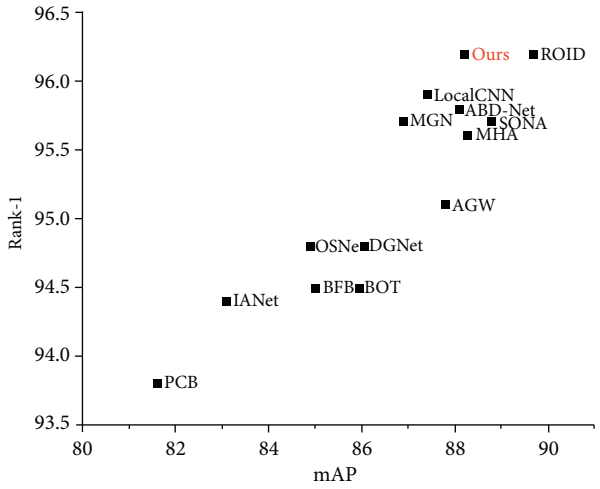

Figure 8: Comparison of indicators between this model and several other Re-ID network models.

Figure 9: Top-10 ranking list for some query images on Market-1501 datasets.

Figure 9. Figure 9 mainly shows the matching results of the images in Ranks-1–10. The green box indicates that the image matches the Gallery library image correctly, belonging to the same ID, and the red box indicates the matching error. It can be seen that, for most pedestrian images, although there are some factors such as background interference, low resolution, and misalignment, the model in this paper can match correctly and achieve high accuracy. It can also be seen from the matching errors of the two red boxes in Figure 9 that the reason for the matching error may be that the backpack covers the important information of the upper body of pedestrians or there are few positive samples, and the search is complete in Rank-8. Therefore, there are many reasons for model matching errors, such as less positive samples, less feature information, and more interference from occlusion and background.

4. Conclusions

Applying attention mechanism to improve the performance of person reidentification model has become a hot topic and has made some progress indeed. However, there are still some problems to be solved, such as matrix operation, coupling problem with original network, and optimizing attention. In this paper, person reidentification model, attention mechanism, and block scheme are studied, and a person reidentification model based on multiattention and multiscale residuals is proposed. Multiple attention models are added to the backbone network and global branches. The multiscale residuals and multiproportion block and reorganization are used to obtain better local and global features. The experimental results show that the model in this paper has some progress in indicators and has certain advantages.

Further, we also notice that although the indicators have improved to some extent, they are still not ideal, and there are still errors in Rank-n. We also analyze the reasons for them. In the next steps, we will continue to study Re-ID problems under occlusion and small samples conditions and try to give improvement or solutions. Another research direction is model lightweight. On the basis of ensuring the recognition rate, the lightweight model can reduce the requirement of system computing power and is better applied to the economic and efficient scene.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

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

References

[1] F. Wang, B. Zhao, C. Huang, and Y. Yan, "Person Re-identification based on multi-scale and attention fusion," *Journal of Electronics and Information Technology*, vol. 42, pp. 1–8, 2020.

[2] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person Re-identification: past, present and future," in *Proceedings of the Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, 2016.

[3] A. Hermans, L. Beyer, and B Leibe, "Defense of the triplet loss for person re-identificationencoding," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1526–1535, Honolulu, HI, USA, 2017.

[4] H. Luo, W. Jiang, X. Fan, and S. Zhang, "A survey on deep learning based person Re-identification," *Acta Automatica Sinica*, vol. 45, no. 11, pp. 2032–2049, 2019.

[5] H. Liu, Z. Jie, K. Jayashree et al., "Video-based person Re-identification with accumulative motion context," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 10, pp. 2788–2802, 2018.

[6] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person Re-identification," in

Proceedings of the Computer Vision and Pattern Recognition, pp. 79–88, Salt Lake City, UT, USA, 2018.

[7] K. Zhou, Y. Yang, A. Cavallaro, and T. Xiang, "Omni-scale feature learning for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3702–3712, Long Beach, CA, USA, 2019.

[8] H. Lawen, A. Bencohen, and M. Protter, "Attention network robustification for person ReID," in *Proceedings of the Computer Vision and Pattern Recognition*, Long Beach, CA, USA, 2019.

[9] R. Hou, B. Ma, H. Chang, X. GU, and S. Shan, "Interaction-and-aggregation network for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 9317–9326, Long Beach, CA, USA, 2019.

[10] B. Chen, W. Deng, and J. Hu, "Mixed high-order attention network for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition Proceedings of the Computer Vision*, pp. 371–381, Long Beach, CA, USA, 2019.

[11] T. Chen, S. Ding, J. Xie, Y. Yuan, and W. Chen, "ABD-net: attentive but Diverse person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 8351–8361, Long Beach, CA, USA, 2019.

[12] X. Bryan, Y. Gong, Y. Zhang, and P. Christian, "Second-order non-local attention networks for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3760–3769, Long Beach, CA, USA, 2019.

[13] X. Zhang, H. Luo, X. Fan et al., "Aligned-reID: surpassing human-level performance in person re-identification," 2018, https://arxiv.org/abs/1711.08184.

[14] H. Zhao, M. Tian, S. Sun, J. Shao, and J. Yan, "Spindle net: person re-identification with human body region guided feature decomposition and fusion," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1077–1085, Honolulu, HI, USA, 2017.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Yian, and S. Wang, "Beyond Part Models: person retrieval with refined Part Pooling (and A strong convolutional baseline)," in *Proceedings of the European conference on computer vision*, pp. 480–496, Munich, Germany, 2018.

[16] G. Wang, Y. Yuan, X. Chen, J. Li, and X. Zhou, *Learning Discriminative Features With Multiple Granularities For Person Re-Identification*, pp. 274–282, ACM multimedia, Chengdu, China, 2018.

[17] F. Zheng, C. Deng, X. Sun, X. Jiang, and R. Ji, "Pyramidal person Re-identification via multi-loss dynamic training," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 8514–8522, Long Beach, CA, USA, 2019.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[19] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-Excitation networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 2011–2023, 2019.

[20] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 7794–7803, Salt Lake City, UT, USA, 2018.

[21] H. Luo, Y. Gu, X. Liao, S. Lai, and W. Jiang, "Bag of tricks and a strong baseline for deep person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 1487–1495, Long Beach, CA, USA, 2019.

[22] D. Fan, S. Fang, G. Wang, S. Gao, and X. Liu, "The visual human face super-resolution reconstruction algorithm based on improved deep residual network," *EURASIP Journal on Advances in Signal Processing*, vol. 1, pp. 1–10, 2019.

[23] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 510–519, Long Beach, CA, USA, 2019.

[24] X. Ma, J. Guo, S. Tang, Z. Qiao, and Q. Chen, "DCANet: learning connected attentions for convolutional neural networks," in *Proceedings of the Computer Vision and Pattern Recognition*, Washington, DC, 2020.

[25] X. Yu, Z. Zhang, L. Wu et al., "Deep ensemble learning for human action recognition in still images," *Complexity*, vol. 2020, no. 1, 23 pages, Article ID 9428612, 2020.

[26] J. Fu, J. Liu, H. Tian, Y. Li, and Y. Bao, "Dual attention network for scene segmentation," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3146–3154, Long Beach, CA, USA, 2019.

[27] G. Wang, S. Gao, and D. Fan, "A G2G similarity guided pedestrian Re-identification algorithm," *Journal of Physics: Conference Series*, vol. 1453, Article ID 012035, 2020.

[28] Y. Lin, W. Chi, W. Sun, S. Liu, and D. Fan, "Human action recognition algorithm based on improved ResNet and skeletal keypoints in single image," *Mathematical Problems in Engineering*, vol. 2020, Article ID 6954174, 12 pages, 2020.

[29] Z. Dai, M. Chen, X. Gu, S. Zhu, and P. Tan, "Batch DropBlock network for person Re-identification and beyond," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3690–3700, Long Beach, CA, USA, 2019.

[30] J. Yang, X. Shen, X. Tian, H. Li, and J. Huang, "Local convolutional neural networks for person reidentifiification," in *Proceedings of the 2018 ACM Multimedia Conference on Multimedia Conference*, pp. 1074–1082, Seoul, Korea, 2018.

[31] X. Qian, Y. Fu, T. Xiang et al., "Pose-normalized image generation for person Re-identification," in *Proceedings of the European conference on computer vision*, pp. 661–678, Munich, Germany, 2018.

[32] Z. Zheng, X. Yang, Z. Yu, L. Zheng, and Y. Yang, "Joint discriminative and generative learning for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 2138–2147, Long Beach, CA, USA, 2019.

[33] Y. Sun, L. Zheng, W. Deng, and S. Wang, "SVDNet for pedestrian retrieval," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3820–3828, Washington, DC, 2020.

[34] M. Ye, J. Shen, G. Lin, T. Xiang, and C. Steven, "Deep learning for person re-identification: a survey and outlook," in *Proceedings of the Computer Vision and Pattern Recognition*, Washington, DC, 2020.

[35] Z. Zhang, C. Lan, W. Zeng, X. Jin, and Z. Chen, "Relation-aware global attention for person Re-identification," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3186–3195, Washington, DC, 2020.

[36] Z. Zhong, L. Zheng, D. Cao, and S. Li, "Re-ranking person Re-identification with k-reciprocal encoding," in *Proceedings of the Computer Vision and Pattern Recognition*, pp. 3652–3661, Honolulu, HI, USA, 2017.

[37] Z. Yang, J. Liu, T. Liu, L. Wang, and S. Zhao, "Circle-based ratio loss for person reidentification," *Complexity*, vol. 2020, no. 12, 11 pages, Article ID 9860562, 2020.

[38] G. Wang, S. Wang, W. Chi, S. Liu, and D. Fan, "A person reidentification algorithm based on improved siamese network and hard sample," *Mathematical Problems in Engineering*, vol. 2020, Article ID 3731848, 11 pages, 2020.