future internet

MDPI

Article

Person Re-Identification Based on Attention Mechanism and Context Information Fusion

Shengbo Chen [1,2], Hongchang Zhang [1] and Zhou Lei [1,*]

[1] School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China; schen@shu.edu.cn (S.C.); hczhang@shu.edu.cn (H.Z.)
[2] Shanghai Key Laboratory of Computer Software Testing and Evaluating, Shanghai 201112, China
* Correspondence: leiz@shu.edu.cn; Tel.: +86-188-0029-5068

Abstract: Person re-identification (ReID) plays a significant role in video surveillance analysis. In the real world, due to illumination, occlusion, and deformation, pedestrian features extraction is the key to person ReID. Considering the shortcomings of existing methods in pedestrian features extraction, a method based on attention mechanism and context information fusion is proposed. A lightweight attention module is introduced into ResNet50 backbone network equipped with a small number of network parameters, which enhance the significant characteristics of person and suppress irrelevant information. Aiming at the problem of person context information loss due to the over depth of the network, a context information fusion module is designed to sample the shallow feature map of pedestrians and cascade with the high-level feature map. In order to improve the robustness, the model is trained by combining the loss of margin sample mining with the loss function of cross entropy. Experiments are carried out on datasets Market1501 and DukeMTMC-reID, our method achieves rank-1 accuracy of 95.9% on the Market1501 dataset, and 90.1% on the DukeMTMC-reID dataset, outperforming the current mainstream method in case of only using global feature.

Keywords: deep learning; person re-identification; attention mechanism; context information fusion; margin sample mining

Citation: Chen, S.; Zhang, H.; Lei, Z. Person Re-Identification Based on Attention Mechanism and Context Information Fusion. *Future Internet* **2021**, *13*, 72. https://doi.org/10.3390/fi13030072

Academic Editor: Remus Brad

Received: 27 January 2021
Accepted: 8 March 2021
Published: 13 March 2021

1. Introduction

The core aim of person re-identification (ReID) is to recognize a person captured in different times and/or locations over multiple non-overlapping camera views, considering a large number of candidates. It has attracted widespread attention in recent years because of its importance in many practical applications, such as image retrieval and video surveillance analysis. The result of person ReID query can not only help the police to quickly obtain clues of criminal suspects or find missing children, but also enable businesses to capture the patterns of customer behavior and allocate resource by making more reasonable use of commercial value [1].

The two key steps in person ReID are pedestrian feature extraction and feature similarity matching, respectively [2]. Traditional person ReID methods mainly depend on the features of manual design and measurement learning. Li et al. [3] propose to overcome the small sample size problem in person ReID distance metric learning, they match people in a discriminative null space of the training data. Liao et al. [4] come up with a method called Local Maximal Occurrence (LOMO), which is helpful for an effective feature representation, and a metric learning method called Cross-view Quadratic Discriminant Analysis (XQDA). The horizontal occurrence of local features is analyzed by LOMO feature, and the occurrence is maximized to make a stable representation against viewpoint changes. Zhao et al. [5] suggest a method of learning mid-level filters from automatically discovered patch clusters for person ReID. The traditional person ReID methods mentioned above often have shortcomings such as insufficient feature representation, low similarity matching accuracy, and slow person ReID process.

The advent of deep learning makes person ReID methods based on deep learning become a hot research topic. There are two mainly methods for person ReID based on deep learning, which are feature representation and deep metric learning, respectively.

- The method based on feature representation is to perform person ReID tasks by learning a network that can extract significant features of pedestrians. Generally, there are three types of features used for person ReID: global feature, local feature, and auxiliary feature. Zheng et al. [6] regards the pedestrian feature extraction as a classification task, and they train the CNN network by using the pedestrian ID as the label of data. Luo et al. [7] proposed a person ReID network using only global features in order to avoid the excessive complexity of network structure. Ye et al. [8] design a new powerful method called AGW (Attention Generalized mean pooling with Weighted triplet loss), which also uses only the global feature of pedestrians, but achieves competitive results. Wu et al. introduced a reverse attention module to handle the problem that the attention mechanism may lead to the loss of important information [9]. Moreover, some researchers are not satisfied with the information contained in global features, so local feature representation learning is proposed. Zhang et al. [10] apply local feature, and address body misalignment problem by design a two-stream network that uses fine grained semantics. To avoid the problem of missing body parts for pedestrians, Fu et al. propose a Horizontal Pyramid Matching (HPM) approach to make full use of pedestrian body parts [11]. To further improve the robustness of person ReID model, some auxiliary information is used to train the model. For example, Attribute Attention NetWork (AANet) is a novel model that combines person attributes with attribute attention maps [12]. In addition to pedestrian attribute information, viewpoint is also a common auxiliary information. Viewpoint-Aware Loss with Angular Regularization (VA-reID) [13] takes into account not only the viewpoint, but also the relationship between different viewpoints. Another popular solution is to combine multiple features for person ReID. Li et al. [14] learn the global and local characteristics of pedestrians by a multi-scale context awareness network (MSCAN) and spatial transformation network (STN). Guo et al. [15] introduce channel and spatial attention mechanism to learn pedestrian characteristics. Harmonious attention network (HAN) [16] framework is designed to solve the high cost of calculation and lengthy inference process, which applies global-local representation learning and combines attention mechanisms.
- The method based on deep metric learning uses CNN to learn the similarity of different pedestrian images, so that the similarity of the same type of pedestrian images is greater than that of different type of pedestrian images. Increasingly, researchers design different kinds of loss function to guide the feature representation learning. Shi et al. [17] train the network by using triplet loss, triplet loss function can make the relative distance of positive sample pairs smaller than that of negative sample pairs. Cheng et al. [18] improve the triplet loss, taking the absolute distance between positive and negative pedestrian samples into account. Xiao et al. [19] present a method of mining difficult samples, using the most dissimilar positive samples and the most similar negative samples to train the network. The combination of multiple loss function is also a common solution. Ye et al. trained visible thermal person ReID model by combing ranking loss and identity loss [20].

The methods based on deep learning have improved speed and accuracy compared with traditional methods, but there are still problems such as inadequate pedestrian features extraction, missing small-scale context information, and complex network structure.

In order to solve the problems of the complex structure of the person ReID network at this stage and the low recognition accuracy due to deformation and occlusion in actual tasks, this paper proposes a person ReID method that only uses global features of pedestrians and introduces an attention mechanism. Using only global features can avoid the complexity of the network structure caused by using several local features. The attention mechanism can control the weight of extracting significant features of pedestrians, thereby improving the semantic information of high-level feature maps. The attention mechanism we use

can increase the accuracy of person ReID while only adding a small amount of network parameters. In addition, in order to solve the problem of the loss of small-scale pedestrian features (such as hats, sunglasses, etc.) due to the increase of the network depth and down-sample, a method that uses different dilated convolution to sample low-level feature maps and concatenate to high-level feature maps is proposed to enhance the representation of pedestrian features. Then, in order to improve the robustness of the network, we combine the Margin Sample Mining Loss (MSML) function and the cross entropy loss function to train the model. Finally, the experimental results on the datasets Market1501 and DukeMTMC-reID show that our method has a certain improvement in identification accuracy.

The contributions of our work can be summarized as follows:

- For the first time, we introduce the ECA attention module in the person ReID task. The ECA attention module increases the accuracy of person ReID tasks while only increasing a few network parameters.
- We design a multi-scale information fusion module, which effectively integrates pedestrian context information and enhances pedestrian feature representation.
- We jointly train the model with MSML and cross entropy loss, so that the robustness of the model is enhanced. Experimental results show that the model we designed performs well on the dataset Market1501 and DukeMTMC-reID, surpasses most mainstream person ReID method in case of only using global feature.

2. Methods

The focus of our method is to learn the significant features of pedestrians. The overall framework of our model is demonstrated in Figure 1. In this section, we introduce our person ReID model from three aspects: the channel attention module for pedestrian salient features learning as shown in Section 2.1, the multi-scale information fusion module for context information extraction (Section 2.2), and the loss function in Section 2.3.

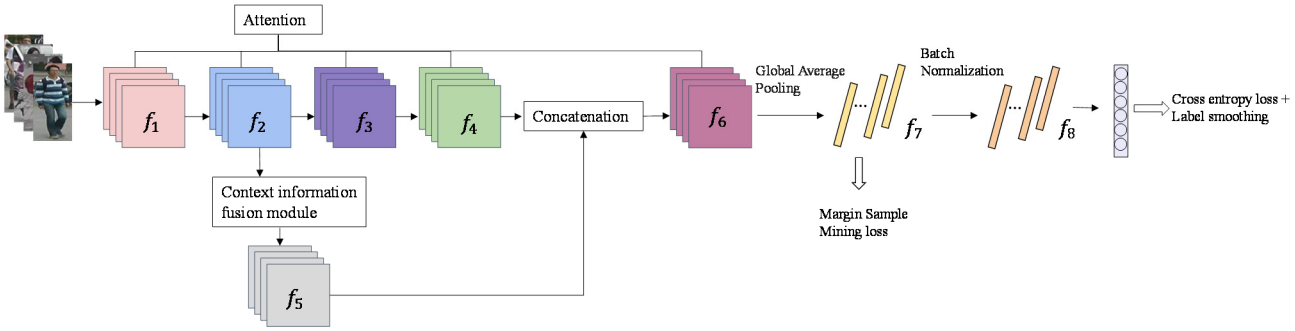


Figure 1. Overall framework of the person re-identificaton (ReID) model we propose. The channel attention module is introduced into the ResNet50. Feature map f_5 is generated by context information fusion module, and it is cascaded with feature map f_4. Feature map f_8 is used for the inference stage.

2.1. Channel Attention Module to Improve the Network

The attention mechanism has been shown to be an effective way to improve deep convolutional neural networks. SE-Net [21] first proposes a method to learn channel attention, and achieves satisfactory results. Then, the development of attention mechanism can be divided into two aspects: (1) the fusion of enhanced features; (2) the integration of channels and spatial attention. CBAM [22] fuses feature by using maximum pooling and average pooling. To achieve more effective feature fusion, a second-order pooling is introduced [23]. GE [24] uses deep convolution to explore spatial expansion to aggregate features. scSE [25] and CBAM use two-dimensional convolution with a kernel size of *k***k* to calculate spatial attention, and then combine spatial attention with channel attention. It can be seen that the above methods obtain effective performance by designing complex

attention modules. Different from the above attention mechanism, the goal of the attention mechanism we used is to learn effective channel attention while reducing the complexity of the model.

Generally, channel attention is achieved by mapping channel features to a low-dimensional space, and then mapping it back, which makes the relationships between channels and their weights indirect. Studies have shown that the change of channel dimensionality and the interaction between channels can affect the performance of channel attention to some extent [26]. Keeping the channel dimension constant helps to learn effective attention, and cross-channel interaction also helps to learn effective attention. Based on the SE module, the Efficient Channel Attention (ECA) module [26] reduces the complexity of the model and improves the efficiency of learning attention by increasing local cross-channel interaction and channel sharing parameters. The channel weight calculation formula is shown as follows:

$$\omega_i = \sigma(\sum_{j=1}^{k} \alpha^j y_i^j), y_i^j \in \Omega_i^k, \tag{1}$$

Ω_i^k represents the k domain channels of y_i, y_i is the feature representation of channel i after global average pooling, α^j is the shared parameter, and σ is the activation function. w_i is the weight of channel i. The structure of the ECA module is shown in Figure 2.

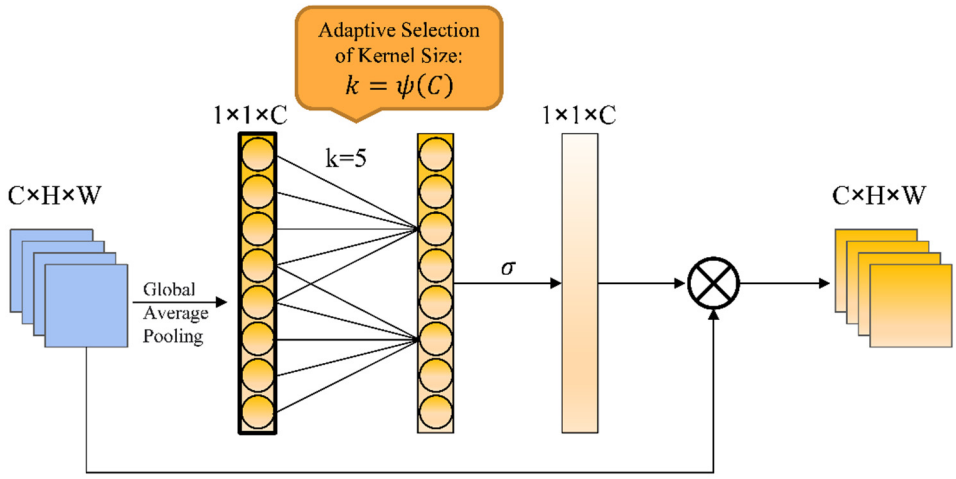


Figure 2. Efficient channel attention (ECA) module diagram. After global average pooling, the channel weights are generated by one-dimensional convolution of size k, and k is adaptively determined by the channel dimension.

The calculation of the weight can be realized by one-dimensional convolution with a kernel size of k. It can be seen that k is a key parameter, because k determines the range of interaction between channels. It is generally believed that the larger the channel dimension, the larger the range of interaction, and the smaller the channel dimension, the smaller the range of interaction. It is reasonable to assume that the interaction range k is proportional to the channel dimension C. A linear function is the simplest mapping, but the relation determined by linear function is too limited. In general, channel dimension C is a power of 2. Thus, this linear relationship can be extended to a nonlinear one, as shown in Formula (2).

$$C = \phi(k) = 2^{(\gamma * k - b)}, \tag{2}$$

Given the channel dimension C, the formula for the one-dimensional convolution kernel size k is shown as follows:

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{odd}, \tag{3}$$

$|t|_{odd}$ is the nearest odd number to t. According to the experiment, setting γ to 2 and b to 1 is more effective.

The improved network with ECA module introduces the channel attention mechanism while only increasing k network parameters. The improved network enhances the semantic information of high-level feature maps when extracting pedestrian features, and suppresses relatively irrelevant feature information, which further improves the robustness of the person ReID feature extraction network. Figure 3 is a partial network structure diagram of the attention module introduced in ResNet50. The feature map is weighted by the attention network after three convolution operations of the backbone network, and then added to the original feature map in the "Eltw sum" method. "Eltw sum" represents the add operation of the feature maps on the corresponding channel. The new feature map is obtained and sent to the next layer through the activation function. In addition, this paper uses the Leaky ReLU activation function to replace the ReLU activation function used in the original ResNet backbone network. The ReLU activation function sets all negative values to zero, which may lead to the loss of some features. A non-zero slope is assigned to all negative values by Leaky ReLU, which is a supplement to the ReLU function and can effectively enhance the extraction of pedestrian features.

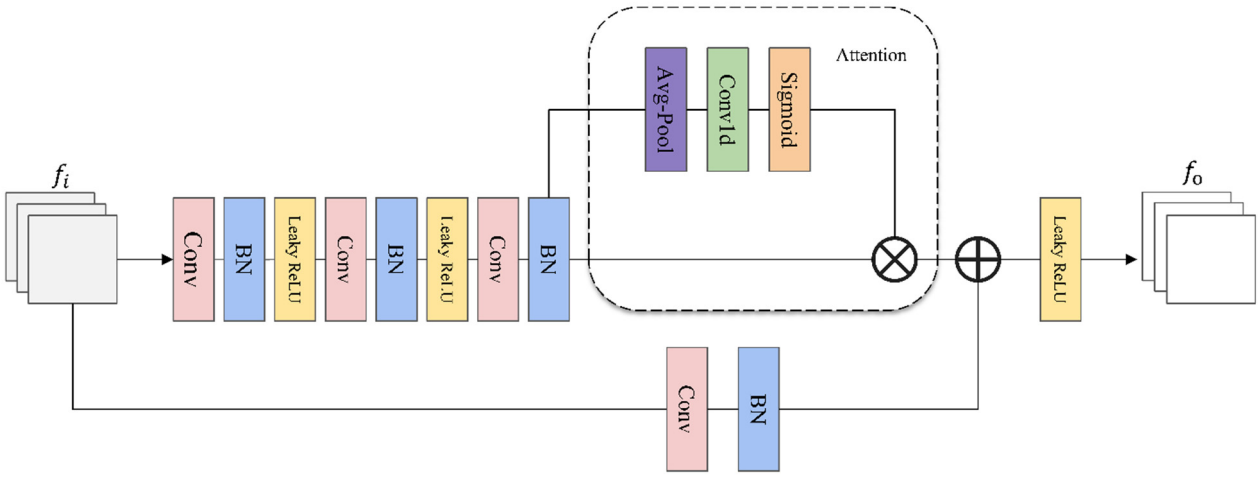


Figure 3. The details of the improved ResNet50 using ECA module.

2.2. Multi-Scale Information Fusion Module

In the process of person ReID, we not only need to extract the high-level features of the pedestrian, but also need to extract the small-scale context information of the pedestrian, such as hats, sunglasses, etc. The deep neural network (DNN) is used to extract the representation of pedestrian features. With the increase of network depth, although high-level features of pedestrians can be obtained, these small features will be lost as the network continuously uses pooling operation to conduct down-sampling of features. Considering this problem, this paper proposes an effective solution that uses convolutions with different dilation ratios to sample shallow feature maps and concatenates them to high-level feature maps, so that we can obtain feature representations containing multi-scale context information of the image. The problem of small-scale information loss due to pooling operations is solved.

The dilated convolutions [27] operation can expand the size of the ordinary convolution kernel according to the dilation ratio. The relationship between the size of the dilated convolution kernel and the size of the original convolution kernel is shown in Formula (4):

$$Filter_2 = d \times (Filter_1 - 1) + 1, \quad (4)$$

where $Filter_2$ represents the size of the convolution kernel after dilation, d represents the dilation ratio, and $Filter_1$ represents the size of the original convolution kernel. It can be seen from Formula (4) that we can use the original convolution of the same size and use different dilation ratios to obtain dilated convolutions of different sizes.

Figure 4 is a schematic diagram of dilated convolution with a core size of 3×3 and dilation ratios of 1, 2, and 3 respectively. Convolution with dilation ratio of 1, 2, and 3 and kernel size of 3×3 is adopted instead of convolution with kernel size of 3×3, 5×5, and 7×7, respectively, because the latter will generate more redundant information and increase computation.

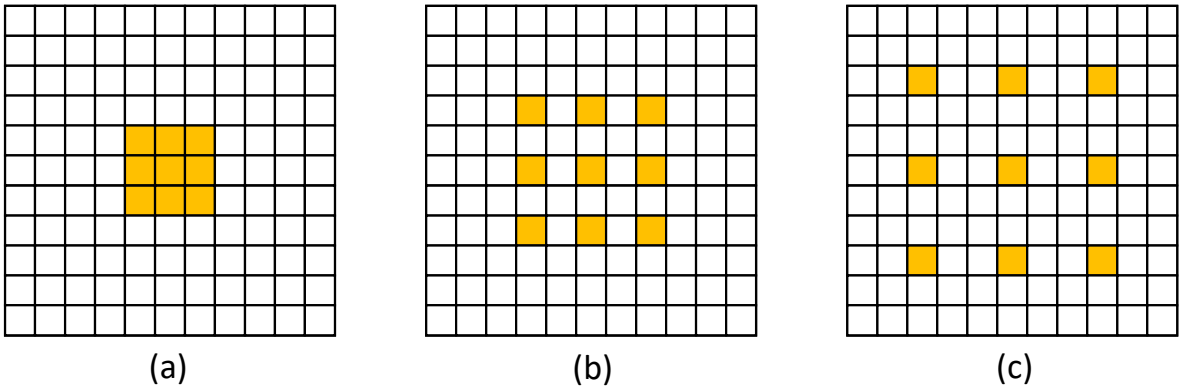


Figure 4. Diagram of dilated convolution. The size of the convolution kernel is 3×3 and the dilation ratio is 1, 2, and 3, respectively. (**a**) Convolution with dilation ratio of 1; (**b**) Convolution with dilation ratio of 2; (**c**) Convolution with dilation ratio of 3.

Inspired by the literature [14], we design a multi-scale information fusion module to extract the context information of pedestrians. As shown in Figure 5, the backbone network extracts shallow features and outputs a feature map with a size of $256 \times 64 \times 32$. We use 3×3 convolutions with the dilation ratios of 1, 2, 3, 4 respectively to sample the feature maps, and then concatenate them together. In order to concatenate the feature map after processing to the high-level feature map output by the backbone network, this paper uses 1×1 convolution to adjust the size of the feature map to $1024 \times 16 \times 8$.

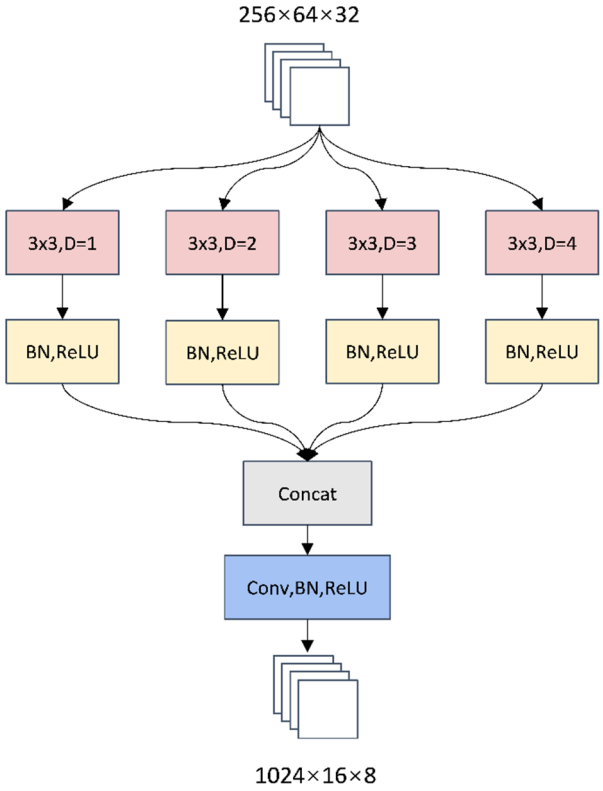


Figure 5. Context Information Fusion Module, where D represents dilation ratio and BN represents batch normalization. The input feature map is operated by using the convolution whose kernel size is 3×3 and the dilation ratios are 1, 2, 3, and 4, respectively.

2.3. Loss Function

2.3.1. Margin Sample Mining Loss

In deep metric learning, the triple loss function is usually used. The triple loss makes the distance of the same type of pedestrians smaller than that of different types to achieve the purpose of clustering the same type of pedestrian images in the feature space. But, the triple loss only measures the relative distance, because the triple loss may not provide a global optimal constraint, so the distance between classes may be smaller than that within the same class. Therefore, the model in this paper adopts the Margin Sample Mining Loss function proposed in [19]. The margin sample mining loss not only considers the absolute distance between positive and negative sample pairs, but also introduces the idea of difficult sample mining. In training, the most dissimilar positive sample pair and the most similar negative sample pair in the whole batch are selected.

$$L_{eml} = (\max_{A,A'}(||f_A - f_{A'}||_2) - \min_{C,B}(||f_C - f_B||_2) + \alpha)_+, \quad (5)$$

In Formula (5), $(x)_+ = \max(x, 0)$. $||f_A - f_{A'}||_2$ represents the Euclidean distance between A and A', and α is a value of the margin used to distinguish positive samples from negative ones. A and A' are the most dissimilar positive sample pairs in each batch of training, and C and B are the most similar negative sample pairs. If A and C are pedestrians in the same category, it considers the relative distance between the positive and negative samples. If A and C are not the same category, it considers the absolute distance between positive and negative samples. For the detailed derivation of this formula, please refer to [19].

2.3.2. Cross Entropy Loss

The cross entropy loss function [7] is used in the classification task of person ReID. The formula of the cross entropy loss function is shown as follows:

$$Lc = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = 0 & y \neq i \\ q_i = 1 & y = i \end{cases}, \quad (6)$$

where N is the number of categories of pedestrians in the training set, y is the real pedestrian label, and p_i is the probability that the network predicts that the pedestrian belongs to label i. In addition, the current person ReID dataset is not very large. To prevent the overfitting phenomenon in the person ReID model training, this paper adopts the label smoothing (LS) [28] operation. LS changes the structure of q_i in Formula (7):

$$q_i = \begin{cases} \frac{\varepsilon}{N} & y \neq i \\ 1 - \frac{N-1}{N}\varepsilon & y = i \end{cases}, \quad (7)$$

where ε is the error rate, which can make the model have better generalization ability. This article sets it to 0.1 according to [7]. N represents the total number of pedestrian categories. When the training data is relatively small, the LS can remarkably improve the performance of the person ReID model. The cross entropy loss function that introduces LS is shown as follows:

$$L_{CL} = \sum_{i=1}^{N} -q_i \log(p_i) \begin{cases} q_i = \frac{\varepsilon}{N} & y \neq i \\ q_i = 1 - \frac{N-1}{N}\varepsilon & y = i \end{cases}, \quad (8)$$

2.3.3. Joint Loss Function

Inspired by literature [7], the person ReID model in this paper is jointly trained with the measurement loss function and the classification loss function. The MSML function can be used to optimize the distance within and between classes, and cross entropy loss with LS is used to predict the probability of pedestrian category. Because the optimization objectives of these two loss functions are inconsistent, in our model, MSML function is used at f_7 in Figure 1 and cross entropy loss with LS is used at f_8. In short, the loss function of our person ReID model is as follows:

$$L_{loss} = L_{eml} + L_{CL}, \quad (9)$$

3. Experiment

3.1. Datasets and Evaluation Metrics

Our experiments are performed on two popular person ReID datasets: Market1501 [29] and DukeMTMC-reID [30]. The partition of the dataset is shown in Table 1. The brief introductions to the datasets are shown below:

Table 1. Details of the datasets used in our experiments.

Dataset	# Image	# ID	# Train	# Validation	# Test
Market1501	32,688	1501	751	751	750
DukeMTMC-reID	36,411	1404	702	702	702

Market1501 is a pedestrian dataset collected at Tsinghua University. Five HD cameras and one ordinary camera captured 1501 pedestrians, and 32,668 detected pedestrian bounding boxes. The picture size is 128 × 56. Among them, there are 751 types of pedestrians in the training set, which contains 12,936 images, that is, an average of 17 pictures per pedestrian; the testing set has 750 types of pedestrians, including 19,732 pictures, that is, an average of 26 pictures per person. In this experiment, we randomly select an image from each type of pedestrian images in the training set, that is, 751 images of 751 types of

pedestrians as a validation set. The remaining 12,185 images in the training set are used as training our model, and the images in the testing set are used to test our model.

DukeMTMC-reID is a pedestrian dataset collected at Duke University in the United States. The pictures are taken by 8 non-overlapping cameras, and the pedestrian frame is manually marked. The training set includes 16,522 pictures, and the testing set includes 17,661 pictures. The training set and the test set include 702 types of pedestrians respectively. In this experiment, we randomly select an image from each type of pedestrian images in the training set, that is, 702 images of 702 types of pedestrians as a validation set. The remaining 15,820 images in the training set are used as training our model, and the testing set images are used to the test model.

The metrics used in this paper to evaluate our model are mAP (Mean Average Precision) and Rank-n.

mAP is the average value of average accuracy (*AP*), that is, the average value of each category of *AP*. To calculate *mAP*, the *AP* is calculated firstly, and *AP* is the area under the PR (Precision Recall) curve. The calculation formulas of precision and recall are as follows, where positive samples with correct predictions is represented as *TP*, positive samples with incorrect predictions are represented as *FP*, and *FN* is negative samples with incorrect predictions.

$$precision = \frac{TP}{TP + FP}, \tag{10}$$

$$recall = \frac{TP}{TP + FN}, \tag{11}$$

AP is the ratio of the sum of all accuracy rates of the category to the number of images in the category, and the calculation formula is as follows (c_i is the number of categories i):

$$AP = \frac{\sum precision}{C_i}, \tag{12}$$

AP measures the effect of a single category, and *mAP* is used to evaluate all categories, as demonstrated in Formula (13), where N is the number of all categories:

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}, \tag{13}$$

Another evaluation index Rank-n represents the probability that the first n query results are correct. For example, Rank-5 represents the probability that the first five images in the query result are correct.

3.2. Experimental Details

Our person ReID models are implemented on PyTorch framework. All experiments are performed under the hardware environment of Intel i9-10900k 3.7GHz CPU and single NVIDIA GeForce RTX3090 GPU.

Before training the model, the image was preprocessed according to [7]. The size of the image was adjusted to 256 × 128 using interpolation method, and pad the image 10 pixels with zero values. Then the image is cropped to 256 × 128 at random. Moreover, each pedestrian image has a 0.5 probability of being flipped horizontally, we also normalize RGB channels by subtracting 0.485, 0.456, and 0.406 and dividing by 0.229, 0.224, and 0.225, respectively. In order to solve the overfitting problem caused by a small scale of data, the method called Random Erasing Augmentation (REA) [31] is also used in our experiment to process the image, the probability of Random Erasing is 0.5. In addition, the batch size is set to 32, the dropout probability is 0.5, and the epoch is 60. Learning rate also plays an important role in the training of person ReID model. We apply a warmup strategy [7] to train our network. Formula (14) shows the relationships between learning rate $lr(t)$ and

epoch t. When the epoch is less than or equal to 10, we make the learning rate increase linearly. At the 40th epoch, the learning rate decays to 0.000035.

$$lr(t) = \begin{cases} 3.5 \times 10^{-5} \times \frac{t}{10} & t \leq 10 \\ 3.5 \times 10^{-4} & 10 < t \leq 40 \\ 3.5 \times 10^{-5} & 40 < t \leq 60 \end{cases} \quad (14)$$

Take the Market1501 dataset as an example, we draw the training and validation loss curve and the top1 error curve. From Figure 6, it can be seen that the training effect is basically optimal when the epoch reaches 40 times. Local optimization has been achieved before the 40th epoch. At the 40th epoch, the learning rate is reduced by 10 times, and the reduced training step size make loss decrease rapidly.

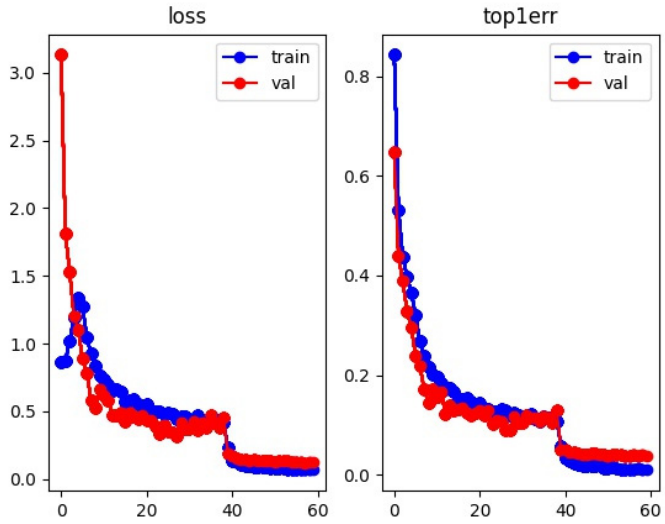


Figure 6. Training and validation loss curve and top1 error curve.

3.3. Experimental Results

After the model is trained, the image features of the query set and the training set are extracted separately, and the Euclidean distance is used to measure the similarity, and the query results are sorted in descending order of probability. Figure 7 shows the query result of the person ReID model in this paper. The ones without borders are the correct pictures, and the ones with black borders are the wrong results.

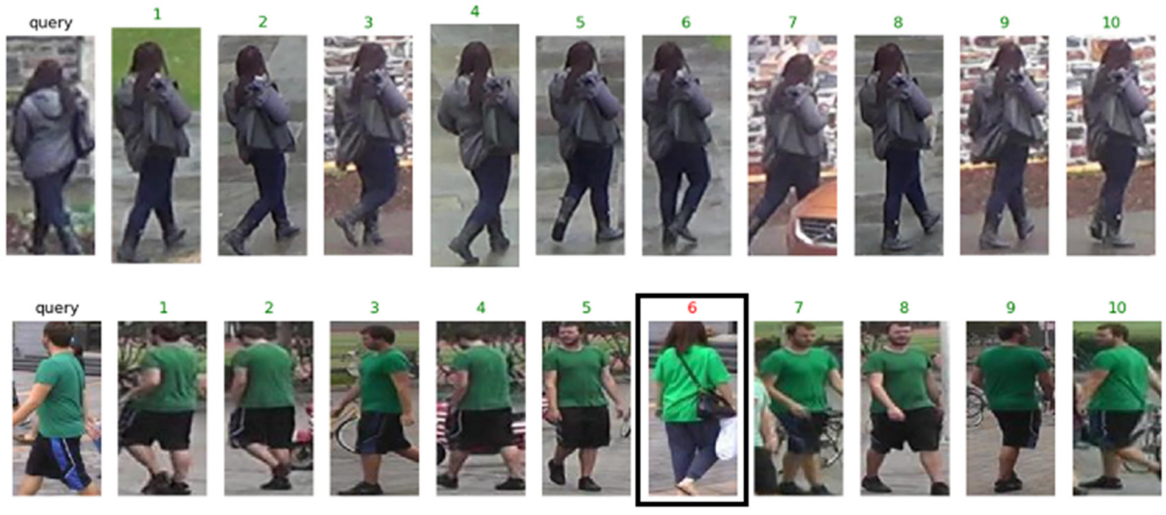


Figure 7. Visualization of person re-identification query results.

3.3.1. Ablation Study

In order to verify the effectiveness of the various modules of the person ReID model in this paper, experiments are carried out on datasets: Market1501 and DukeMTMC-reID. We sequentially add modules for model training. The Baseline is a model with ResNet50 as the backbone network and the loss function as cross entropy. L1 represents the attention module, L2 represents the context information fusion module, and L3 represents the combination of MSML and cross entropy loss. Tables 2 and 3 show the results of the ablation experiment. Rank-1 and mAP increased by 1.3% and 2.9%, respectively, on Market1501 and increased by 2.7% and 3.6%, respectively, on DukeMTMC-reID after adding the attention module into the baseline. Furthermore, rank-1 and mAP increased by 1.4% and 1.4%, respectively, on Market1501 and increased by 1.1% and 2.8%, respectively, on DukeMTMC-reID after adding attention mechanism and context information fusion module. Finally, on the basis of adding attention module and context information fusion module, the MSML and cross entropy loss are used to jointly train the model. On Market1501 dataset, Rank-1 and mAP reach 95.2% and 87.1%, respectively; on DukeMTMC-reID dataset, Rank-1 and mAP reach 88.9% and 78.5%, respectively. Experimental results show that the person ReID model proposed in this paper enhances the representation of pedestrian features and improves the accuracy of identification.

Table 2. Comparison of ablation experiment results (Market1501).

	Baseline	**Baseline + L1**	**Baseline + L1 + L2**	**Baseline + L1 + L2 + L3**
Rank-1	91.8	93.1	94.5	95.2
mAP	81.8	84.7	86.1	87.1

Table 3. Comparison of ablation experiment results (DukeMTMC-reID).

	Baseline	**Baseline + L1**	**Baseline + L1 + L2**	**Baseline + L1 + L2 + L3**
Rank-1	82.6	85.3	86.4	88.9
mAP	70.5	74.1	76.9	78.5

3.3.2. Comparison with Mainstream Methods

In order to verify the superiority of the model in this paper, we also compare with the current mainstream methods on the datasets Market1501 and DukeMTMC-reID. We divide these mainstream person ReID methods into six categories: traditional methods, deep metric learning, local feature-based methods, global-local based feature methods, auxiliary feature-based methods, and global feature-based methods. It can be seen from Tables 4 and 5 that the experimental results of our method show better performance on the two datasets. DSA [10] achieves the surprising performance on Market1501 dataset without using re-ranking. However, this method uses a set of local features of the pedestrian, and multi-branch network structure makes the model too complicated. Without re-ranking, the performance of the ADMS [9] method using global features on the two datasets is slightly higher than our method, but ADMS architecture includes five branches and five loss functions, and our model only has two branches with two loss functions. After adding re-ranking [32] to optimize, the rank-1 and mAP of our model on the Market1501 dataset are increased to 95.9% and 94.5%, respectively. On the DukeMTMC-reID dataset, the rank-1 and mAP are increased to 90.1% and 89.6%, respectively. The results of experiment show that our method outperforms the current mainstream method in case of only using global feature.

Table 4. Comparison of the results of different methods under the Market1501 dataset.

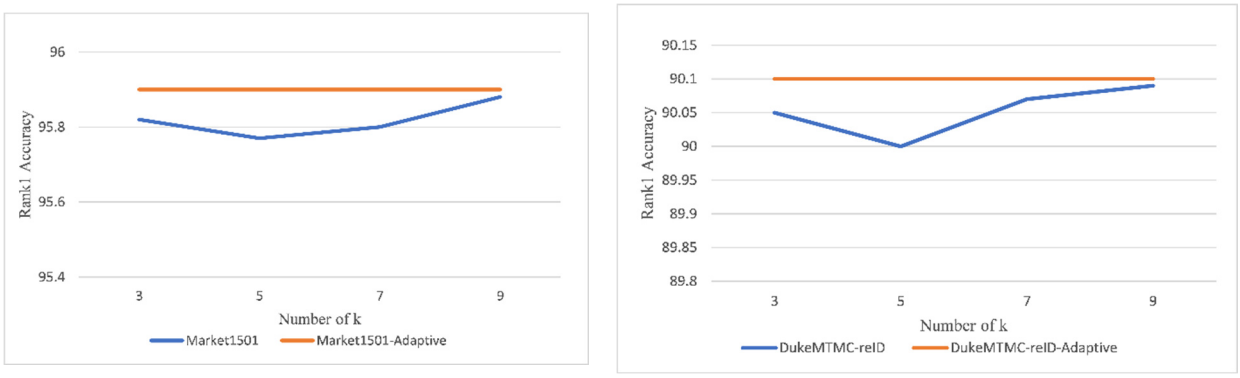
Type	Method	Rank-1	mAP
Traditional methods	XQDA [4]	43.0	21.7
	NPD [3]	55.4	30.0
Deep metric learning	Quad [33]	80.0	61.1
	TriNet [17]	84.9	69.1
	MSML [19]	85.2	69.6
Local feature	PCB [34]	92.3	77.4
	PAN [35]	81.0	63.4
	DSA [10]	95.7	87.6
	HPM [11]	94.2	82.7
Global-local feature	MSCAN [14]	80.3	57.5
	HAN [16]	93.1	89.6
Auxiliary feature	AACN [36]	85.9	66.9
	AANet [12]	93.9	83.4
Global feature	IDE [6]	81.9	61.0
	Mancs [37]	93.1	82.3
	BagTricks [7]	94.5	85.9
	ADMS [9]	95.5	89.0
	Ours	95.2	87.1
	Ours + Re-ranking	95.9	94.5

Table 5. Comparison of the results of different methods under the DukeMTMC-reID dataset.

Type	Method	Rank-1	mAP
Traditional approaches	XQDA [4]	31.2	17.2
	NPD [3]	46.7	27.3
Deep metric learning	Quad [33]	73.4	58.0
Local feature	PCB [34]	81.7	66.1
	PAN [35]	71.6	51.5
	DSA [10]	86.2	74.3
	HPM [11]	86.6	74.3
Global-local feature	HAN [16]	84.6	81.3
Auxiliary feature	AACN [36]	76.8	59.3
	AANet [12]	87.6	74.2
Global feature	Mancs [37]	84.9	71.8
	BagTricks [7]	86.4	76.4
	ADMS [9]	89.4	79.2
	Ours	88.9	78.5
	Ours + Re-ranking	90.1	89.6

3.3.3. Effect of Kernel Size (k) on Our Model

According to the analysis in Section 2.1 of this paper, k in Formula (1) and (2) is the key parameter. In practice, k is also the size of 1D convolution kernel. In this section, its effects on our person ReID model are evaluated. We train our model by setting k be 3 to 9. The results of the experiment can be obtained from Figure 8. First of all, we fix the value of k over all the convolution blocks, and our model achieves the best results when k = 9. Additionally, when we adopt the adaptive strategy of Formula (2), the results outperform the fixed ones. The above analyses show the effectiveness of the adaptive strategy in Formula (2) in obtaining stable and better results.



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

Figure 8. The effects of different k values on the model ((**a**): Market1501 dataset, (**b**): DukeMTMC-reID). Also, we give the results when the kernel size k is selected adaptively.

4. Conclusions

In order to improve the shortcomings of existing person ReID methods in pedestrian features extraction, we propose a person ReID model based on attention mechanism and context information fusion. The effective channel attention can enhance the salient features of pedestrians and suppress irrelevant features. In addition, we design a context information fusion module which can help us to address the problem of small-scale context information loss. Finally, to further improve robustness, we combine MSML with cross entropy loss function to train the model. The experimental validation and analysis show that the method in this paper is robust and has good performance on the two mainstream datasets. The future work is to further optimize the model to improve the accuracy of person ReID without increasing the complexity of the network.

Author Contributions: Conceptualization, H.Z.; Formal analysis, H.Z.; Project administration, Z.L.; Software, H.Z.; Validation, S.C. and Z.L.; Writing—original draft, H.Z.; Writing—review & editing, S.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61572306 and Grant 61502294, in part by the National Key Research and Development Program of China under Grant 2017YFB0701600, and in part by the Shanghai Engineering Research Center of Intelligent Computing System under Grant 19DZ2252600.

Data Availability Statement: Data is contained within the article.

Acknowledgments: The authors thank the High-Performance Computing Center of Shanghai University for providing computing resources.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wang, H.; Du, H.; Zhao, Y.; Yan, J. A comprehensive overview of person re-identification approaches. *IEEE Access* **2020**, *8*, 45556–45583. [CrossRef]
2. Jiang, M.; Li, Z.; Chen, J. Person Re-Identification Using Color Features and CNN Features. In Proceedings of the 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, 20 March 2019; pp. 460–462.
3. Zhang, L.; Xiang, T.; Gong, S. Learning a discriminative null space for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27 June 2016; pp. 1239–1248.
4. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 12 June 2015; pp. 2197–2206.
5. Zhao, R.; Ouyang, W.; Wang, X. Learning mid-level filters for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 28 June 2014; pp. 144–151.
6. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
7. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [CrossRef]

8. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *arXiv* **2021**, arXiv:2001.04193.
9. Wu, D.; Wang, C.; Wu, Y.; Wang, Q.C.; Huang, D.S. Attention deep model with multi-scale deep supervision for person re-identification. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *5*, 70–78. [CrossRef]
10. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Densely semantically aligned person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 667–676.
11. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Huang, T. Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 1 February 2019; Volume 33, pp. 8295–8302.
12. Tay, C.P.; Roy, S.; Yap, K.H. Aanet: Attribute attention network for person re-identifications. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7134–7143.
13. Zhu, Z.; Jiang, X.; Zheng, F.; Guo, X.; Huang, F.; Sun, X.; Zheng, W. Aware Loss with Angular Regularization for Person Re-Identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–8 February 2020; Volume 34, pp. 13114–13121.
14. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
15. Guo, T.; Wang, D.; Jiang, Z.; Men, A.; Zhou, Y. Deep network with spatial and channel attention for person re-identification. In Proceedings of the 2018 IEEE Visual Communications and Image Processing (VCIP), Taichung, Taiwan, 9–12 December 2018; pp. 1–4.
16. Li, W.; Zhu, X.; Gong, S. Scalable Person Re-Identification by Harmonious Attention. Available online: https://link.springer.com/content/pdf/10.1007/s11263-019-01274-1.pdf (accessed on 8 March 2021).
17. Shi, H.; Yang, Y.; Zhu, X.; Liao, S.; Lei, Z.; Zheng, W.; Li, S.Z. Embedding deep metric for person re-identification: A study against large variations. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 732–748.
18. Cheng, D.; Gong, Y.; Zhou, S.; Wang, J.; Zheng, N. 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, Las Vegas, NV, USA, 30 June 2016; pp. 1335–1344.
19. Xiao, Q.; Luo, H.; Zhang, C. Margin sample mining loss: A deep learning based method for person re-identification. *arXiv* **2017**, arXiv:1710.00478.
20. Ye, M.; Lan, X.; Wang, Z.; Yuen, P.C. Bi-directional center-constrained top-ranking for visible thermal person re-identification. *IEEE Trans. Inf. Forensics Secur.* **2019**, *15*, 407–419. [CrossRef]
21. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 7132–7141.
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Con-ference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
23. Gao, Z.; Xie, J.; Wang, Q.; Li, P. Global second-order pooling convolutional networks. In Proceedings of the IEEE/CVF Con-ference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3024–3033.
24. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Vedaldi, A. Gather-excite: Exploiting feature context in convolutional neural networks. *arXiv* **2018**, arXiv:1810.12348.
25. Roy, A.G.; Navab, N.; Wachinger, C. Recalibrating fully convolutional networks with spatial and channel "squeeze and excitation" blocks. *IEEE Trans. Med Imaging* **2018**, *38*, 540–549. [CrossRef] [PubMed]
26. Qilong, W.; Banggu, W.; Pengfei, Z.; Peihua, L.; Wangmeng, Z.; Qinghua, H. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11534–11542.
27. Yu, F.; Koltun, V. Multi-scale context aggregation by dilated convolutions. *arXiv* **2015**, arXiv:1511.07122.
28. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 30 June 2016; pp. 2818–2826.
29. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; Tian, Q. Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 13 December 2015; pp. 1116–1124.
30. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 17–35.
31. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–8 February 2020; Volume 34, pp. 13001–13008.
32. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking person re-identification with k-reciprocal encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 26 July 2017; pp. 1318–1327.
33. Chen, W.; Chen, X.; Zhang, J.; Huang, K. Beyond triplet loss: A deep quadruplet network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 26 July 2017; pp. 403–412.

34. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 480–496.
35. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3219–3228.
36. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 23 June 2018; pp. 2119–2128.
37. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 365–381.