electronics

Article

Weakly Supervised Cross-Domain Person Re-Identification Algorithm Based on Small Sample Learning

Huiping Li [1], Yan Wang [2], Lingwei Zhu [3], Wenchao Wang [1], Kangning Yin [2,3,4], Ye Li [3,4] and Guangqiang Yin [2,*]

[1] School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China; hpli@std.uestc.edu.cn (H.L.); wenchao_wang@std.uestc.edu.cn (W.W.)
[2] School of Information and Software Engineering, University of Electronic Science and Technology of China, Chengdu 610051, China; lywang@std.uestc.edu.cn (Y.W.); knyin@std.uestc.edu.cn (K.Y.)
[3] Shenzhen Institute for Advanced Study, University of Electronic Science and Technology of China, Shenzhen 518110, China; zhulingwei@std.uestc.edu.cn (L.Z.); liyeye@std.uestc.edu.cn (Y.L.)
[4] Institute of Public Security, Kash Institute of Electronics and Information Industry, Kash 844000, China
* Correspondence: yingq@uestc.edu.cn

Abstract: This paper proposes a weakly supervised cross-domain person re-identification (Re-ID) method based on small sample data. In order to reduce the cost of data collection and annotation, the model design focuses on extracting and abstracting the information contained in the data under limited conditions. In this paper, we focus on the problems of strong data dependence, weak cross-domain capability and low accuracy in Re-ID in weakly supervised scenarios. Our contributions are as follows: first, we implement a joint training framework with the help of small sample learning and cross-domain migration for Re-ID. Second, with the help of residual compensation and fusion attention module, the RCFA module is designed, and the model framework is built on this basis to improve the cross-domain ability of the model. Third, to solve the problem of low accuracy caused by insufficient data coverage of small samples, a fusion of shallow features and deep features is designed to enable the model to weighted fusion of shallow detail information and deep semantic information. Finally, by selecting different camera images in Market1501 dataset and DukeMTMC-reID dataset as small samples, respectively, and introducing another dataset data for joint training, we demonstrate the feasibility of this joint training framework, which can perform weakly supervised cross-domain Re-ID based on small sample data.

Keywords: person re-identification; weakly supervision; small sample; cross-domain migration

Citation: Li, H.; Wang, Y.; Zhu, L.; Wang, W.; Yin, K.; Li, Y.; Yin, G. Weakly Supervised Cross-Domain Person Re-Identification Algorithm Based on Small Sample Learning. *Electronics* **2023**, *12*, 4186. https://doi.org/10.3390/electronics12194186

Academic Editor: Donghyeon Cho

Received: 1 September 2023
Revised: 26 September 2023
Accepted: 3 October 2023
Published: 9 October 2023

1. Introduction

Person re-identification (Re-ID) [1] refers to the association and matching of a specific target person using computer vision techniques in scenarios across devices, times and locations. The task is generally viewed as a fine-grained image retrieval problem with constraints. Re-ID can make up for the face recognition technology and fixed camera vision limitations in the fields of intelligent security and video surveillance; and can be combined with person detection and person tracking [2] for Re-ID systems. Traditional video surveillance systems, due to their low degree of intelligence, may result in criminal investigators not only needing to consume a lot of time and energy, but also inevitably negligent omissions on the way to work. In addition, the suspect may confuse the criminal investigator by deliberately covering the face or wearing different color clothing to achieve the purpose of obstructing the tracking. The emergence of Re-ID technology can, to a certain extent, solve the problem of inefficiency and the high rate of missed detection that exists in traditional video surveillance. With the construction of smart cities, security needs are increasing day by day, and intelligent monitoring systems have ushered in a major development opportunity. As an indispensable part of it, Re-ID has become a hot research direction in the academic and industrial fields.

With the improvement of computer hardware performance such as GPU and the huge amount of data brought by the big data environment, Re-ID algorithms based on deep learning have developed rapidly. The Re-ID method based on deep learning integrates two modules: feature extraction and metric learning, that is, the extraction of image features and the similarity comparison of feature vectors are completed in one model. According to the different Re-ID methods, the deep learning-based Re-ID models can be divided into representation models [3] and matching models [4], where the representation models treat the Re-ID task as a classification problem and the loss functions of the representation models are classification loss [5] and verification loss [3], etc.

Typically, fully supervised deep learning models require large amounts of labeled data for training, and expecting data to cover all sample characteristics is alien to the idea of data-driven model learning. Therefore, model design should focus less on the use of large amounts of labeled data and more on extracting and abstracting the information and knowledge contained in the data under limited data conditions. This leads to the concept of weakly supervised learning. Weakly supervised learning of small amounts of labeled data in a weakly supervised scenario is of great value and significance for the implementation of applications related to Re-ID systems. The cross-domain migration-based approach is unsupervised learning by domain adaptation, where the model is pre-trained in a supervised manner on labeled source data and then adapted to the target domain of unlabeled data. Through migration learning, the domain knowledge of the labeled dataset is transferred to the unlabeled dataset. However, existing Re-ID techniques mostly rely on supervised learning with a large number of samples, which requires a large amount of labeled data for training, and thus cannot be generalized to other scenarios, requiring labor costs and computational costs that are not conducive to the implementation of Re-ID techniques. Therefore, weakly supervised Re-ID techniques based on small sample learning are needed to achieve cross-domain Re-ID.

In this paper, we focus on the problems of strong data dependence, weakly cross-domain capability and low accuracy encountered in Re-ID in weakly supervised scenarios, and implement a joint training framework based on deep learning with the help of small-sample learning and cross-domain migration for Re-ID. The framework uses single-camera labeled data as a small-sample training set and introduces a large amount of data from non-target domains as prior knowledge to improve the Re-ID performance of the model through joint training. The main work of this paper is as follows:

1. To solve the Re-ID problem in weakly supervised scenes, a joint training framework combining cross-domain migration learning and small-sample learning is proposed, which can train both small-sample data and different-domain data to reduce the realistic scene data collection efforts and yet ensure that the algorithm models are adequately trained.
2. To solve the problem of weak cross-domain capability, a Re-ID module residual compensation and fusion attention (RCFA) based on residual compensation and fusion attention is designed. In order to effectively utilize the introduced non-target domain data and solve the disturbance caused by different data distribution, RCFA module is designed, which can suppress inter-domain differences.
3. To further improve the accuracy of the algorithm, a fusion of shallow and deep features is designed so that the model can weight fuse shallow detail information and deep semantic information to solve the model learning bias caused by insufficient coverage of small sample data.

The rest of this article is organized as follows. Section 2 introduces the related work of cross-domain Re-ID and weakly supervised Re-ID. Section 3 introduces our proposed small sample learning weakly supervised cross-domain Re-ID algorithm, and proves the research results of this paper through the experiments in Section 4. Finally, in Section 5, we give our conclusion.

2. Related Work

2.1. Cross-Domain Re-ID

The purpose of cross-domain is to adapt the source domain with fully labeled samples to the target domain with sparse labels. The existing domain adaptation methods are divided into supervised, unsupervised and weakly supervised. Weak supervision only requires image-level annotation, and achieves a balance between the adaptation effect and annotation cost. The purpose of cross-domain weakly supervised object detection is to adjust the object-level knowledge of the fully labeled source domain dataset to train the object detector of the weakly labeled target domain [6]. Therefore, this paper used weak supervision technology to realize the adaptation of the detector from the source domain to the target domain. ICCM [7] divided each image into semantic clusters and aligns the foreground region in the target domain with the labeled region in the source domain. Ref. [8] proposed a cross-domain weakly supervised object detection (CDWSOD) method based on DETR, which aims to adapt the detector from the source domain to the target domain through weakly supervision. H2FA R-CNN [9] enforced two image-level alignments on the backbone features, and performs two instance-level alignments on the RPN and the detection head, effectively narrowing the gap between the source domain and the target domain.

Image style transformation is a transfer learning method in the image domain, first proposed by Gatys in [10]. Because it can effectively solve the problem of model generalization due to image style differences, researchers have widely applied it to cross-domain Re-ID tasks. For example, Deng et al. [11] proposed SPGAN, an unsupervised domain adaptive framework consisting of SiaNet and CycleGAN [12]. The samples generated by coordination between SiaNet and CycleGAN not only have the style of the target domain, but also retain the underlying identity information. An instance-guided context presentation method is proposed in [13], which transfers the source domain person identity to a different target domain context in order to achieve supervised Re-ID in the unlabeled target domain. Yc et al. [14] proposed a new style migration framework STReID, which can change the style while preserving the image content information, and then use both the original image and Zhu et al. [15] decomposed person images into foreground, background and style features, and then use these features to synthesize person images with target domain background for training. In addition to this, there are many studies in improving the generalization ability of the model. Ref. [16] proposed a domain-invariant mapping network (DIMN) to learn the mapping between images and classifiers. It followed a meta-learning pipeline and samples a subset of the source domain training task in each training set to make the model domain invariant.

2.2. Weakly Supervised Re-ID

Based on the problems of supervised and unsupervised Re-ID, researchers have introduced the concept of weakly supervised learning, which is a combination of supervised and unsupervised learning methods to train effective Re-ID models using only a small amount of data. With the increasing interest in weakly supervised learning, a large number of related research branches have emerged and been attributed to it, such that weakly supervised learning has become a comprehensive research field that covers a variety of studies that attempt to build predictive models with weak supervision [17]. Weakly supervised learning, as the name implies, refers to the lack of adequate supervision of the data provided, and from this perspective, semi-supervised learning can be seen as the first and most fundamental framework in the field [18].

Semi-supervised learning (SSL) aims to use both labeled and unlabeled data to accomplish a specific learning task. The concept of semi-supervised learning first appeared in [19]. As the earliest semi-supervised methods, self-learning methods are considered as an iterative mechanism that uses initially labeled data to train a model to predict some unlabeled samples. Then, the most plausible predictions are marked as the best predictions for the current model, thus providing more training data for the supervised algorithm.

The joint training approach [20] provided a similar solution by training two different models on two different views and using the reliable predictions from one view as labels for the other model. Figueira et al. [21] proposed a multi-class learning approach. Given any set of features, regardless of their number, dimensionality and descriptors, the method works by fusing these features and ensuring that they are consistent with the classification results. Li et al. [22] proposed a new semi-supervised region metric learning method that learns discriminative region-to-point metrics by estimating positive neighborhoods to generate positive regions.

Unsupervised learning does not require labeled data and is therefore more adaptive and robust. Early unsupervised Re-ID mainly learns invariant components, i.e., dictionary learning [23], metric learning [24] or significance analysis [25], which leads to limited discriminability or scalability.

Ye et al. [26] proposed an unsupervised cross-camera label estimation method to build a sample map for each camera, iteratively update the label estimation and sample map, and implement cross-camera label association using a dynamic graph matching (DGM) method to solve the problem of poor quality of feature representation and noise generated by cross-views during the association process. Wang et al. [27] proposed a consistent cross-view matching (CCM) framework using global camera network constraints to ensure the consistency of matching pairs, and a cross-view matching strategy using global camera network constraints to explore the matching relationships across the camera network and solve the problem of inaccurate matching results for different camera pairs.

For end-to-end unsupervised Re-ID, Fan et al. [28] first pseudo-labeled the target domain in a cross-domain dataset and proposed an iterative clustering model for Re-ID, first training a convolutional network on the source domain, then going to the target domain for image feature extraction, clustering by K-Means to a set number of families, fine-tuning the model with the clustered results, and so on iteratively. The pseudo-label clustering algorithm combining hierarchical clustering and hard-batch triplet loss proposed by Zeng et al. [29] made full use of the similarity between samples in the target dataset through hierarchical clustering, and reduced the influence of difficult samples through hard-batch triplet loss, resulting in high-quality pseudo-labels and improving model performance. The TAUDL proposed by Li et al. [30] trained an end-to-end neural network by using unsupervised single-camera trajectory information, and then used this image model to automatically label and learn cross-camera images. Most unsupervised learning does not consider the distribution differences between cameras. Xuan et al. [31] iteratively optimized the similarity between cameras by generating pseudo labels within and between cameras.

In recent years, the performance of Re-ID algorithms based on weakly supervised methods has been significantly improved, but there is still a big gap compared with methods based on supervised learning. At present, there are relatively few studies on weakly supervised Re-ID algorithms in academia, and the development is not yet complete. How to transfer the knowledge learned from labeled source datasets to unlabeled target datasets through a domain adaptive approach to achieve higher performance of weakly supervised algorithms will be the focus of related research.

2.3. Small Sample Learning

Small sample learning aims to learn and generalize models through a small number of samples. In practical application scenarios, there are small samples or small labeled samples, and labeling a large number of unlabeled samples will consume a lot of manpower. So, people are committed to studying a small number of samples for learning. Ref. [32] proposed a new small sample target detection algorithm, which does not require fine-tuning and can directly perform small sample target detection on unknown classes. LSTD [33] proposed a regularized transfer learning framework, which can flexibly transfer from the source domain to the target domain, avoiding task differences. Ref. [34] proposed a new adversarial learning method to learn feature representation, which can not only achieve domain adaptation, but also achieve class separation in a specific domain.

Nakamura et al. [35] proposed a fine-tuning method, which uses a lower learning rate in the process of retraining on small sample categories and can be achieved by adjusting the entire network when there is a big difference between the source dataset and the target dataset. In the field of Re-ID, limited by privacy protection laws, it is difficult to obtain people monitoring data, so it is necessary to adopt small sample learning.

3. The Proposed Method

3.1. Model Basic Framework

In order to solve the problem of insufficient training data for small sample learning, prior knowledge must be introduced to assist model learning, and different types of prior knowledge have different effects. If a large amount of labeled data different from the target domain is introduced, the amount of data for the overall training of the model can be expanded. However, there are differences between different domains, and directly introducing them as training data into training does not guarantee a positive effect. Therefore, this paper designs a basic framework, which can ensure that the introduced prior knowledge has a positive effect. By enhancing the model's ability to extract invariant features, it can effectively utilize a large amount of labeled data in the non-target domain. The overall structure is shown in Figure 1. The framework uses ResNet-50 [36] as the backbone network and inserts RCFA modules at different stages to extract domain invariant features.

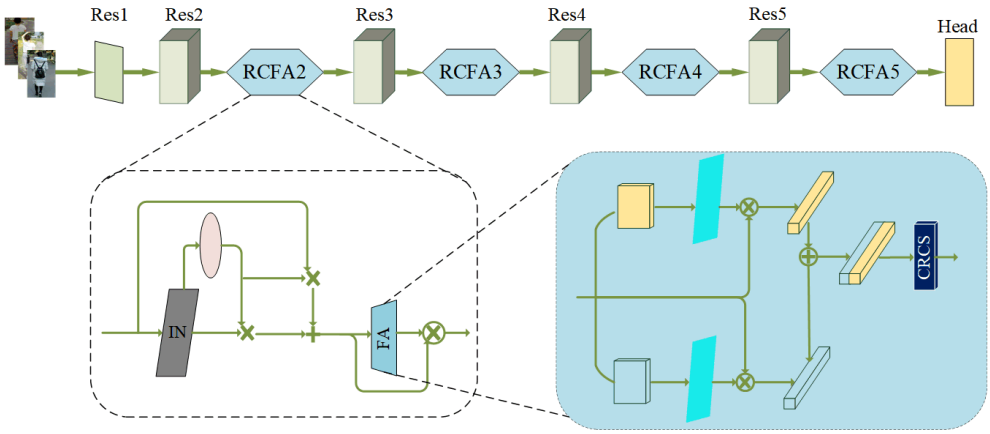

Figure 1. Model basic framework.

3.2. RCFA Module

In order to improve the generalization ability of the Re-ID model, we design a residual compensation and fusion attention (RCFA) module based on instance normalization (IN) residual connection structure and fusion attention (FA) module (as shown in Figure 2).

The FA module helps the model to extract more discriminative semantic features at the cost of adding a small amount of computation. C, H and W denote the number of channels, height and width of the features, respectively. In order to introduce spatial information into the channel dimension, the spatial information map needs to be obtained first. For the input feature $F \in R_c \times h \times w$, Average Pooling and Max Pooling operations are performed along the channel dimension to aggregate the channel features, and two two-dimensional maps are obtained. Average Pooling can limit the variance of the estimate due to the restricted neighborhood size by selecting the average pixel value of a certain region to represent the overall features of the region. Max Pooling is used to select the maximum pixel value in a region to represent the overall features in that region, which helps to retain the saliency information in the feature map and gives the model some resistance to distortion. Second, to efficiently compute the final fused attention weights, the spatial dimensions of the F_{avg} and F_{max} maps need to be compressed. Therefore, the feature maps with fused spatial information then performed Average Pooling and Max Pooling operations along the spatial axis to aggregate spatial information and generate two

spatial contextual descriptors, respectively. The computational results focus on the global information and saliency information in the feature maps that help to discriminate people, respectively. Finally, the two spatial context descriptors are combined and fused by a CRCS (Conv + ReLu + Conv + Sigm) module to obtain the final fused attention weights.

In particular, to suppress the effect of inter-domain differences, the IN normalized data distribution is used to suppress style differences. And the input features x and the normalized features x_{IN} are fused by residual concatenation to compensate for the person discriminative information lost during the instance normalization calculation. Then, the person features are further enhanced by calculating weights and weighting them by the FA module. Figure 3 shows the structure of the RCFA module.

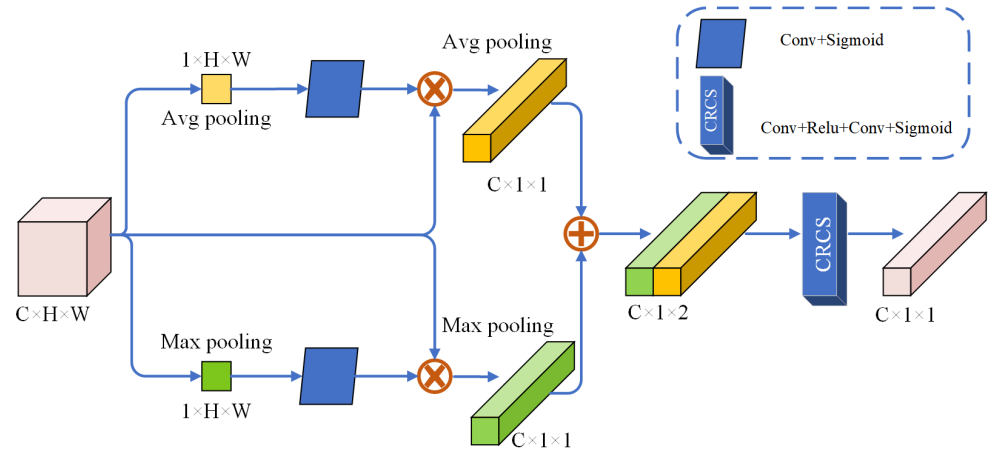

Figure 2. FA module.

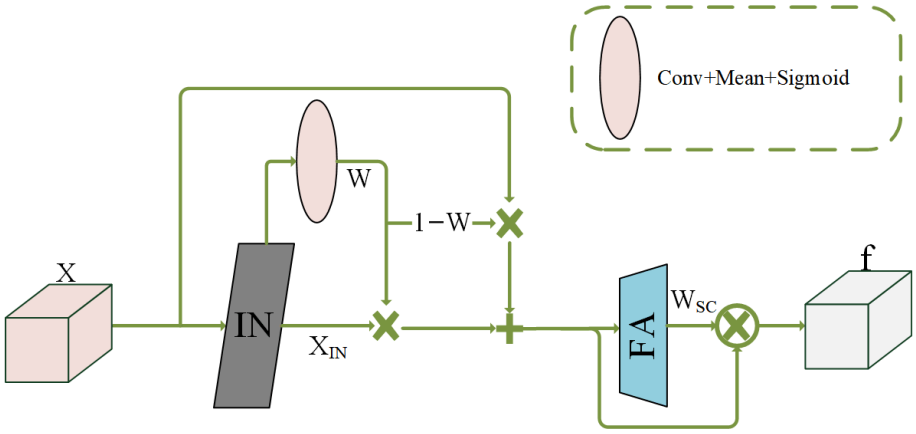

Figure 3. RCFA module.

The information $x \in R^{c \times h \times w}$ carried by the input features includes style (activated mean and variance) and shape (activated spatial structure) information. Instance normalization suppresses the style differences between different domains by calculating the mean and variance in each channel of each sample while the shape information remains unchanged. The calculation is as follows:

$$x_{IN} = IN(x) = \gamma \times \frac{x - \mu(x)}{s(x)} + \beta \tag{1}$$

where, $\mu(\cdot)$ and $s(\cdot)$, respectively, represent the average and standard deviation calculated in the space size of each channel. γ and β are the parameters learned by the network model during training.

3.3. Joint Training Framework

In order to solve the problem of Re-ID in weakly supervised scenarios, a joint training framework combining cross-domain transfer learning and small sample learning is proposed. Aiming at the problem of building the Re-ID algorithm model in the real environment, cross-domain transfer learning and small sample learning can effectively reduce the data collection work and calculation cost, and ensure that the model is adequately trained. In a large number of Re-ID studies, due to the complete training data, the model can learn more sufficient features from a variety of different resolution samples. For the Re-ID model, sufficient training data with different resolutions is crucial to improve its generalization ability. For each image in the training set, if all the corresponding images with the same content but different resolutions can be obtained, it will help the model to obtain better generalization ability. However, a small amount of learning methods are used to reduce the collection of real scene data. The selected method is to select a camera sample for labeling. At this time, while greatly reducing the amount of data collection and labeling, it also greatly reduces the number of samples with different resolutions, making it difficult to extract sufficient information in the model learning process. In order to solve the above problems, this part further improves the basic framework to form the final proposed joint training framework, as shown in Figure 4.

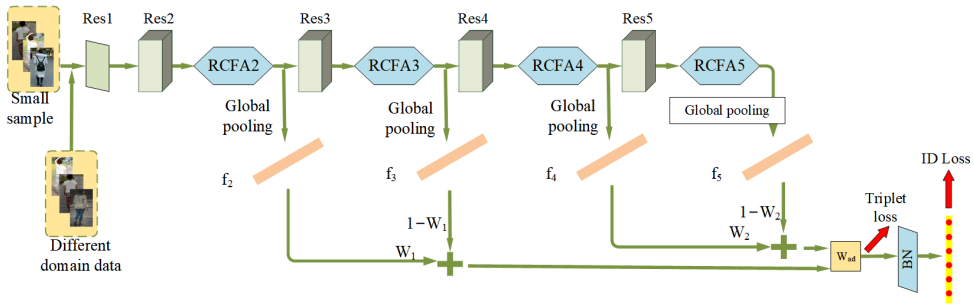

Figure 4. Joint training framework after feature fusion.

The specific feature fusion process is as follows: Firstly, we input the feature maps output after each RCFA module of the basic framework into the global pooling module to obtain features f_2, f_3, f_4 and f_5. Features f_2 and f_3 contain similar information, while features f_4 and f_5 contain similar information. In the feature division, it is considered that features f_2 and f_3 mainly contain shallow detail information, while f_4 and f_5 mainly contain deep semantic information. In this paper, f_2 and f_3 are weighted and fused according to the weight w_1, f_4 and f_5 are weighted and fused according to the weight w_2, and the formulas are as follows:

$$f_{23} = w_1 \times f_2 + (1 - w_1) \times f_3 \tag{2}$$

$$f_{45} = w_2 \times f_4 + (1 - w_2) \times f_5 \tag{3}$$

where, the weights w_1 and w_2 are calculated by the triplet loss value. The idea is that the triplet loss function [37] reflects the closeness of the anchor sample and the positive sample in the feature space, and the distance between the anchor sample and the negative sample in the feature space. Therefore, the triplet loss can judge whether the information contained in the feature is sufficiently discriminative. The weights of w_1 and w_2 are calculated according to Formulas (4) and (5):

$$w_1 = e^{-0.7 \times \left(\frac{L_{trip2} + \delta}{L_{trip3} + \delta} \right)} \tag{4}$$

$$w_2 = e^{-0.7 \times \left(\frac{L_{trip4} + \delta}{L_{trip5} + \delta} \right)} \tag{5}$$

where, δ ensures that the denominator will not be zero. In the formula, the ratio of the loss value is scaled by 0.7 times to ensure that the weight given is close to 0.5 when the loss values of the two features are the same. By using the negative exponential power function

of e instead of linear mapping, it is assumed that a larger weight will be given only when the molecule is much smaller than the denominator.

After obtaining the shallow fusion feature f_{23} and the deep fusion feature f_{45}, the two need to be weighted to obtain the fusion feature that finally contains shallow information and deep information. The calculation method is as follows:

$$f_{fusion} = w_{ad} \times f_{23} + (1 - w_{ad}) \times f_{45} \tag{6}$$

where, the weight w_{ad} is an adaptive learning weight to ensure that the model adaptively allocates the proportion of shallow information and deep information in the fusion process. Finally, the fusion feature is passed through the BNNeck structure [38], and the triplet loss value and ID classification loss are calculated, respectively, to constrain the model convergence process.

3.4. Loss Function

This paper uses crossentropy (CE) loss and triplet loss [37] to jointly constrain model training. However, in general, CE loss and triplet loss have inconsistent embedding spaces in the constrained optimization process, and it is easy to have one loss value decrease and another loss value increase [39]. Therefore, this paper introduces the BNNeck [38] structure to improve the model based framework, as shown in Figure 5.

The purpose of Re-ID is to match all images belonging to the same ID as the query image from the image library. In terms of the nature of the task, it can still be divided into an image classification task, in which each person ID is taken as a class and the ID number is labeled. For image classification problems, CE loss is the most commonly used loss function, which has two common types, binary classification and multi-classification. The binary classification represents that there are only two types of prediction results after training the model. In Re-ID, the prediction is expressed as positive samples and negative samples. The formula is as follows:

$$L_{bce} = \frac{1}{N}\sum_{i} L_i = \frac{1}{N}\sum_{i} -[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)] \tag{7}$$

where, y_i represents the label of sample i, the positive sample is 1, and the negative sample is 0; p_i represents the probability that sample i is predicted to be a positive sample. The probability of this. Multi-classification is an extension of binary classification. In Re-ID, each row ID is treated as a class, and the formula is as follows:

$$L_{ce} = \frac{1}{N}\sum_{i} L_i = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic} \cdot \log(p_{ic}) \tag{8}$$

where, M denotes the number of classes, that is, the number of IDs; y_{ic} means that when the real class of sample i is c, take 1, otherwise take 0; p_{ic} represents the probability that sample i is predicted to be a class c.

The core idea of the triplet loss function is to shorten the distance of samples with the same ID in the vector space as much as possible, and to push the distance between samples with different IDs farther. Because its idea is concise and clear and fully conforms to the logical thinking of Re-ID, triplet loss has become a commonly used loss function in the research of Re-ID algorithms. The formula is as follows:

$$L_{trip} = \max(d(a, p) - d(a, n) + margin, 0) \tag{9}$$

In the formula, a, p, n represent the anchor sample, positive sample and negative sample, respectively; the function d is often a euclidean distance metric function; margin is a hyperparameter that can be initially set.

We introduce the BNNeck [38] structure on the basis of the model framework, as shown in Figure 5. The structure adds a BN layer between the final stage of feature extraction and the fully connected layer of the classifier, and initializes the BN layer and the fully connected layer. In the forward propagation stage, Feature f is input into triplet loss to calculate the loss, and then Feature f is input into the BN layer to obtain Feature f_{BN}, which is classified using a fully connected layer. Finally, the probability of ID classification is output to calculate CE loss.

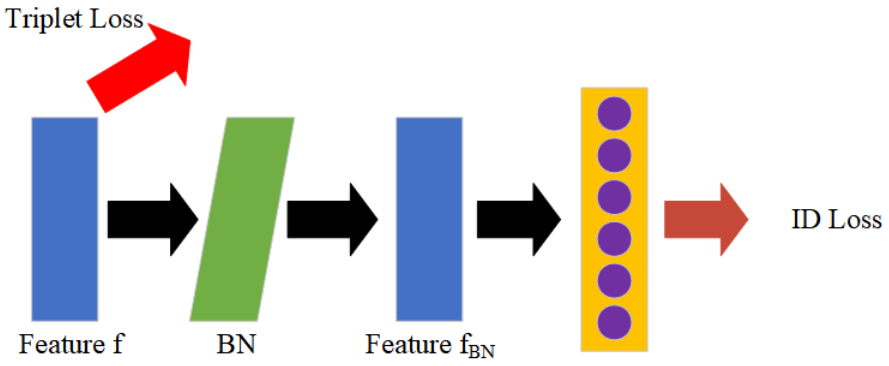

Figure 5. Loss function.

4. Experiment

4.1. Experimental Preparation

The experimental environment is shown in Table 1.

Table 1. Experimental environment.

Category	Type
CPU	Intel(R) Xeon(R) CPU E5-267
RAM	128 G
Hard Disk	8 T
GPU	2 × Nvidia TITAN RTX 24 G
Operating System	Ubuntu 18.04
CUDA	11.0
Python	3.6.9
Pytorch	1.5.1

Four datasets are selected for the experiments in this paper: Market1501 [40], DukeMTMC-reID [41], CUHK03-NP [42] and the large-scale dataset MSMT17 [43].

Market1501 was collected by five HD cameras and one regular camera on the Tsinghua University campus during the summer. The people in the dataset are divided into 751 training IDs and 750 query IDs.

DukeMTMC-reID contains 36,411 images of 1812 people captured by eight HD cameras. Overall, 702 IDs were randomly selected from the dataset and the corresponding 16,522 images were used as the training set, and 2228 images from the remaining 702 IDs were used as the query images. CUHK03-NP is a new training and test set splitting protocol of CUHK03, which splits the training set and test set into 767 and 700 IDs. MSMT17 is collected under different time periods and weather conditions and contains 126,441 labeled borders with 4,101 IDs. Among them, 32,621 labeled borders of 1041 IDs are training sets; the 93,820 labeled borders of 3060 IDs are the test set. The details of each dataset are shown in Table 2.

Table 2. Distribution of datasets.

Datasets	Cameras	Training IDs	Training Images	Test IDs	Test Images	Query Images
Market1501	6	751	12,936	750	19,732	3368
DukeMTMC-reID	8	702	16,522	702	17,661	2228
CUHK03-NP	10	767	7365	700	5332	1400
MSMT17	15	1041	32,621	3060	93,820	11,659

4.2. Experiment Setting

First, data preprocessing is performed by scaling all images to 256 × 128 and padding them with 10 px, performing random horizontal flipping and re-random clipping to 256 × 128. In addition, data augmentation methods such as random color dithering and random patching are used to enhance the diversity of the samples. In the training phase, 4 people with 6 images each are randomly selected from the training set to obtain a small batch size of 24 to train the model. The initial learning rate is set to 3.5×10^{-6}, and the learning rate is increased to 3.5×10^{-4} by Warmup at the 2000th iteration, and then the cosine annealing mechanism is started at the 8000th iteration to continuously reduce the learning rate to 7.7×10^{-4}. Different training iterations are set on different datasets to train the model until convergence.

4.3. Evaluation Indicators

To accurately evaluate the model performance, Rank-1 matching rate and mean average precision (mAP) are used as evaluation metrics. The calculations are as follows:

In conducting the experimental tests, the person features of query and gallery are compared using the cosine distance with the following equation:

$$dist = 1 - cos(Q, G) \tag{10}$$

where, Q and G are both feature vectors (in this article, their dimensions are 1×2048), Q is the normalized query person feature vector and G is the normalized gallery person feature vector. After obtaining the cosine distance, it provides the basis for the subsequent evaluation metrics calculation.

In this paper, two evaluation metrics, mAP (mean Average Precision) and Rank-1, are adopted in the experimental process. When the number of images searched by the model in the dataset is X, only x images out of X are actually the same person to be detected. The model accuracy can be calculated as

$$Precision = \frac{x}{X} \tag{11}$$

The average accuracy (AP) of the person can be further calculated as

$$AP = \frac{\sum Precision}{n} \tag{12}$$

Finally, the average accuracy of all the different types of people is then averaged to mAP:

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

Rank-1 indicates the probability that the first retrieval result is the correct result among the search results returned according to the similarity level for all samples to be tested.

4.4. Ablation Experiment

4.4.1. RCFA Effect Validation

This section sets up ablation experiments to verify the effectiveness of each component of the RCFA module in a cross-domain scenario. Market1501, DukeMTMC-reID and CUHK03-NP are used as target domains, while other datasets from non-target domains are used as source domains to simulate different cross-domain scenarios. In the experiments, ResNet-50 is used as the baseline network; "+RES" indicates the addition of IN-based residual connectivity structure to ResNet-50; "+RCFA" represents the further addition of FA module to the previous one.

As shown in Tables 3–5, when the RCFA module is inserted into the model, the best results are achieved in all cross-domain Re-ID scenarios. Overall, the cross-domain Re-ID performance of CUHK03-NP is poor compared to other target domains. This is because CUHK03-NP is far away from other domains in the feature space. By adding each component one by one and conducting experiments to evaluate the performance of each component, it can be found that each component effectively improves the cross-domain Re-ID performance.

Table 3. Results of the ablation experimental data when Market1501 is the target domain.

Method	Source: CUHK03-NP (%)				Source: DukeMTMC-reID (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	28.4	55.7	71.7	78.0	28.6	57.2	73.3	79.5	31.7	60.5	76.9	83.4
+RES	33.4	62.3	77.1	82.7	33.8	65.1	79.9	85.4	38.0	66.9	81.9	86.8
+RCFA	34.5	63.3	78.8	83.9	34.2	65.8	81.0	86.2	38.9	68.9	82.7	87.0

Table 4. Results of the ablation experimental data when DukeMTMC-reID is the target domain.

Method	Source: CUHK03-NP (%)				Source: Market1501 (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	14.6	30.1	45.8	52.4	24.6	43.4	58.6	64.7	39.6	59.2	75.7	81.0
+RES	20.9	40.1	56.0	62.1	28.8	49.6	65.6	70.6	45.4	65.9	78.1	82.4
+RCFA	21.4	41.0	56.7	63.7	30.9	53.5	67.0	72.3	46.0	67.0	79.4	83.3

Table 5. Results of the ablation experimental data when CUHK03-NP is the target domain.

Method	Source: Market1501 (%)				Source: DukeMTMC-reID (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	14.2	13.6	27.4	36.0	8.7	8.8	18.5	25.0	14.7	14.7	26.8	34.2
+RES	17.0	17.5	33.0	42.0	9.8	10.4	21.4	28.2	16.3	15.7	30.6	39.9
+RCFA	17.7	18.5	33.6	43.4	10.7	11.9	24.4	31.2	19.0	19.4	34.4	44.3

For different cross-domain scenarios, the IN-based residual connection structure is first embedded into the baseline, and it can be seen that the Re-ID effect has been significantly improved. In the DukeMTMC-reID to CUHK03-NP cross-domain scenario, the improvement of mAP metric is the weakest at 1.1%, while the improvement of mAP metric is the most significant at 6.3% for the MSMT17 to Market1501 and CUHK03-NP to DukeMTMC-reID cross-domain scenarios. Other metrics also improved significantly, indicating that the residual linkage effectively normalizes the style of features and preserves discriminative information. Then, the FA module is further introduced to form the complete RCFA module. Similarly, the performance of Re-ID in most cross-domain scenarios is further improved. The experiments demonstrate that the introduction of the RCFA module improves the model cross-domain effect significantly.

Non-local (NL) [44], squeeze and excitation (SE) [45] and convolutional block attention module (CBAM) [46] are three common attention implementation methods in the field of computer vision. In the experiment, ResNet-50 is still used as the baseline network, and the NL, SE, and CBAM structures are inserted into the network, respectively. Market1501, DukeMTMC-reID, and CUHK03-NP are used as target domains, and other data sets of non-target domains are used as source domains to simulate different cross-domain scenarios. The experimental results are shown in Tables 6–8. In most cross-domain scenarios, the Re-ID effect of RCFA is better than NL, SE and CBAM. It can be seen from these tables that adding any attention mechanism to the benchmark network can improve the cross-domain Re-ID performance of the model, which proves that the attention mechanism can also improve the feature extraction ability of the model in cross-domain scenarios, optimize the performance of the model, and enhance generalization ability of the model.

Table 6. RCFA and other network comparison results when the target domain is Market1501.

Method	Source: CUHK03-NP (%)		Source: DukeMTMC-reID (%)		Source: MSMT17 (%)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ResNet-50	28.4	55.7	28.6	57.2	31.7	60.5
ResNet-50 + NL	30.0	57.2	34.9	65.6	38.0	67.4
ResNet-50 + SE	29.4	56.5	32.0	62.7	38.2	67.1
ResNet-50 + CBAM	31.2	61.8	33.8	65.7	38.2	68.4
ResNet-50 + RCFA	**34.5**	**63.3**	34.2	**65.8**	**38.9**	**68.9**

Table 7. RCFA and other network comparison results when the target domain is DukeMTMC-reID.

Method	Source: CUHK03-NP (%)		Source: Market1501 (%)		Source: MSMT17 (%)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ResNet-50	14.6	30.1	24.6	43.4	39.6	59.2
ResNet-50 + NL	14.4	28.0	30.3	52.9	**46.3**	66.1
ResNet-50 + SE	14.1	27.6	25.5	45.9	44.4	65.7
ResNet-50 + CBAM	21.1	38.3	30.4	52.7	45.3	66.0
ResNet-50 + RCFA	**21.1**	**41.0**	**30.5**	**53.5**	45.7	**66.2**

Table 8. RCFA and other network comparison results when the target domain is CUHK03-NP.

Method	Source: Market1501 (%)		Source: DukeMTMC-reID (%)		Source: MSMT17 (%)	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
ResNet-50	14.2	13.6	8.7	8.8	14.7	14.7
ResNet-50 + NL	16.1	17.1	9.2	9.7	16.5	l7.0
ResNet-50 + SE	14.1	14.1	8.3	8.6	14.2	l5.4
ResNet-50 + CBAM	16.9	16.9	9.9	11.3	18.2	l8.4
ResNet-50 + RCFA	**17.7**	**18.5**	**10.7**	**11.9**	**19.0**	**19.4**

Compared to SE, NL is generally in a leading position in terms of performance. Similarly, for CBAM, except for the cross-domain scenarios of DukeMTMC-reID to Market1501, MSMT17 to Market1501, and MSMT17 to DukeMTMC-reID, CBAM is fully ahead of NL and SE in other situations, demonstrating the effectiveness of CBAM as a lightweight insertion module in cross-domain scenarios. RCFA achieves excellent results in almost all cross-domain scenarios. RCFA is only slightly inferior to NL in the cross-domain scenario from MSMT17 to DukeMTMC-reID, with a slight disadvantage of 0.6% in the mAP, while there is no significant difference in the Rank-1.

Under the cross-domain problem, different Re-ID datasets are different. For example, Market-1501 is collected on the domestic campus in the summer, and the people are short-sleeved and the colors are relatively bright; DukeMTMC-reID is collected on winter foreign campuses, and the people wear heavy winter clothes and their colors are relatively dark. This significant difference in dress style shapes the domain gap between the two data sets and reduces the accuracy of the model's cross-domain use between the two datasets, which puts forward high requirements for the generalization ability of the model. Our RCFA performs well on cross-domain problems, so it also performs well when facing overfitting.

4.4.2. Exploration of the Insertion Position of RCFA

In order to discover the effect of different insertion positions of RCFA modules on the network, the framework uses ResNet-50 as the backbone network and inserts RCFA modules at different stages to extract domain invariant features. To assess the impact of different insertion positions of RCFA modules on the generalization ability of the model, experiments are designed to compare various different architectures. Market1501, DukeMTMC-reID and CUHK03-NP are used as the target domains, while other datasets that are not the target domains are used as the source domains to simulate different cross-domain scenarios. Meanwhile, ResNet-50 is set as the Baseline network, and Baseline-RCFA2, Baseline-RCFA3, Baseline-RCFA4, and Baseline-RCFA5 are set to represent the insertion of RCFA modules after different residual blocks in the skeleton of ResNet-50, as shown in Figure 1. While Baseline-RCFA23 represents inserting RCFA modules after residual block 2 and residual block 3 at the same time, Baseline-RCFA2345 represents adding RCFA modules after all residual blocks in the network, and so on.

Tables 9–11 show the performance comparisons of several different architectures in different cross-domain scenarios, where red color indicates the highest value in the vertical comparison, while blue color indicates the second highest in the vertical comparison. It is clear that adding RCFA modules after each residual block of the ResNet-50 network can effectively improve the cross-domain generalization performance of the model compared to Baseline. The optimal case of a single insertion of the RCFA module is demonstrated, as shown in Table 12. It is obvious that Baseline-RCFA4 completely outperforms the other architectures in most scenarios. The performance of Baseline-RCFA4 and Baseline-RCFA5 is comparable in the cross-domain scenarios of DukeMTMC-reID to Market1501 and MSMT17 to DukeMTMC-reID. When CUHK03-NP is not considered, both Baseline-RCFA4 and Baseline-RCFA5 outperform Baseline-RCFA2. The semantic information contained in the deep features is more concentrated in the channel dimension, which is more conducive to the utilization of the RCFA module. However, when considering CUHK03-NP, the performance of Baseline-RCFA5 changes, and it tends to not be as good as Baseline-RCFA2 or Baseline-RCFA3 for model enhancement. This may be due to the fact that the erroneous person detection in CUHK03-NP allows RCFA5 to assist the model in learning erroneous semantic information at deep features.

We further explore the impact of different insertion combinations on the model, as shown in Figure 6, which shows that Baseline-RCFA23 significantly underperforms the other combinations in most of the cross-domain scenarios, suggesting that the RCFA module is more effective in processing deep semantic features. The exception is that Baseline-RCFA45 underperforms Baseline-RCFA23 when extended from DukeMTMC-reID to CUHK03-NP. This suggests that although the RCFA module is good at handling deep features, in some scenarios, enhancing shallow features may be more beneficial for Re-ID. The performance of Baseline-RCFA23, Baseline-RCFA234, and Baseline-RCFA2345 is improved sequentially, which suggests that adding the RCFA module at the deep layer can effectively improve the generalization ability of the model. Baseline-RCFA45, Baseline-RCFA345, and Baseline-RCFA2345 achieve the best results in different cross-domain scenarios, indicating that inserting RCFA modules in the shallow layer does not always improve the generalization ability of the model when RCFA modules are already inserted in the deep layer of the network. Overall, the performance of Baseline-RCFA2345 is more stable than

the other two architectures, thus suggesting that joint processing of shallow and deep features is more effective in improving the model generalization ability.

Table 9. Experimental data results of different architectures when Market1501 is used as the target domain.

Method	Source: CUHK03-NP (%)				Source: DukeMTMC-reID (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	28.4	55.7	71.7	78.0	28.6	57.2	73.3	79.5	31.7	60.5	76.9	83.4
Baseline-RCFA2	28.5	56.5	72.9	78.7	30.0	61.8	77.6	83.1	32.8	62.3	78.3	84.1
Baseline-RCFA3	30.7	59.0	74.6	80.0	31.3	62.7	78.0	83.7	34.3	65.0	79.3	84.8
Baseline-RCFA4	31.6	59.5	75.4	82.1	32.6	63.4	78.9	84.3	37.5	66.7	82.0	86.4
Baseline-RCFA5	28.6	56.0	72.7	78.1	32.2	62.9	78.6	84.4	34.8	65.3	79.1	84.5
Baseline-RCFA23	31.5	60.2	76.0	81.6	29.7	61.6	77.2	83.1	33.4	64.5	78.2	83.8
Baseline-RCFA234	32.4	61.0	77.6	83.1	33.1	64.4	80.3	85.1	37.3	67.9	81.3	86.0
Baseline-RCFA345	33.9	61.7	78.2	83.3	32.9	63.9	79.4	84.8	37.8	67.4	81.5	85.9
Baseline-RCFA45	32.4	59.1	75.2	80.9	34.2	66.1	80.4	85.5	39.3	68.7	82.0	86.8
Baseline-RCFA2345	34.5	63.3	78.8	83.9	34.2	65.8	81.0	86.2	38.9	68.9	82.7	87.0

Red color indicates the highest value in the vertical comparison, while blue color indicates the second highest in the vertical comparison.

Table 10. Experimental data results of different architectures when DukeMTMC-reID is used as the target domain.

Method	Source: CUHK03-NP (%)				Source: Market1501 (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	14.6	30.1	45.8	52.4	24.6	43.4	58.6	64.7	39.6	59.2	75.7	81.0
Baseline-RCFA2	17.6	36.1	51.0	57.8	27.7	48.5	63.7	69.6	43.2	64.3	76.7	81.4
Baseline-RCFA3	17.0	34.6	49.6	57.0	28.4	50.0	64.6	70.3	42.1	62.7	76.8	81.3
Baseline-RCFA4	18.4	34.8	51.2	57.3	28.2	47.8	64.2	68.9	44.1	64.9	77.0	81.3
Baseline-RCFA5	15.1	30.8	46.2	52.9	29.6	51.2	66.7	71.9	44.3	65.5	78.4	82.1
Baseline-RCFA23	17.0	34.7	50.8	56.4	27.3	48.7	63.8	69.3	42.7	63.6	76.9	81.4
Baseline-RCFA234	19.4	37.7	54.6	60.6	29.1	49.6	65.0	70.6	44.7	65.4	78.4	82.1
Baseline-RCFA345	20.5	39.7	55.6	62.2	31.1	53.6	67.2	71.4	44.5	64.1	77.8	81.9
Baseline-RCFA45	21.6	41.1	55.1	62.2	30.6	52.7	66.7	72.1	45.2	65.4	78.0	82.9
Baseline-RCFA2345	21.4	41.0	56.7	63.7	30.9	53.3	67.0	72.3	46.0	67.0	79.4	83.3

Red color indicates the highest value in the vertical comparison, while blue color indicates the second highest in the vertical comparison.

Table 11. Experimental data results of different architectures when CUHK03-NP is used as the target domain.

Method	Source: Market1501 (%)				Source: DukeMTMC (%)				Source: MSMT17 (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
Baseline	14.2	13.6	27.4	36.0	8.7	8.8	18.5	25.0	14.7	14.7	26.8	34.2
Baseline-RCFA2	15.5	16.6	30.0	38.5	9.5	9.8	21.0	27.3	15.5	14.4	28.3	37.5
Baseline-RCFA3	14.9	15.1	29.5	38.4	9.5	10.4	21.8	28.4	15.1	14.8	27.6	35.6
Baseline-RCFA4	15.6	16.0	29.9	38.2	10.5	11.4	22.6	29.9	16.6	17.9	30.5	38.1
Baseline-RCFA5	15.6	16.2	29.4	37.4	9.3	9.8	20.5	27.9	15.4	14.9	27.7	35.4
Baseline-RCFA23	16.0	16.6	32.5	40.4	9.8	10.4	21.5	27.0	15.7	15.8	29.0	38.7
Baseline-RCFA234	17.6	17.6	33.6	42.0	9.9	10.4	21.4	28.2	16.3	16.4	31.9	40.6
Baseline-RCFA345	17.1	17.5	33.5	42.1	10.2	10.6	22.4	30.2	16.9	17.1	30.0	38.1
Baseline-RCFA45	17.4	18.1	32.1	40.6	9.7	9.9	20.4	27.1	17.4	17.4	30.9	39.5
Baseline-RCFA2345	17.7	18.5	33.6	43.4	10.7	11.9	24.4	31.2	19.0	19.4	34.4	44.3

Red color indicates the highest value in the vertical comparison, while blue color indicates the second highest in the vertical comparison.

Table 12. Optimization of single insertion of RCFA modules.

C → M RCFA4	D → M RCFA4, RCFA5	MS → M RCFA4
C → D RCFA4	M → D RCFA5	MS → D RCFA4, RCFA5
M → C RCFA2, RCFA4, RCFA5	D → C RCFA4	MS → C RCFA4

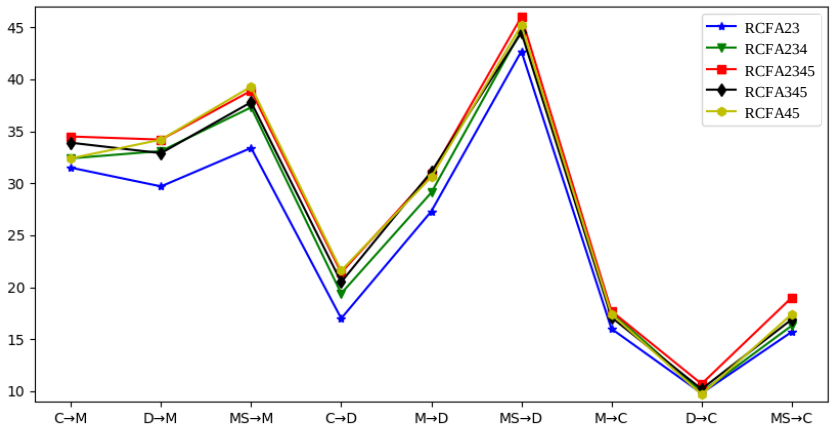

Figure 6. The mAP visualization of different combinations of insertion methods.

4.4.3. Joint Training

This section focuses on the single-camera annotation experiments for Market1501 and DukeMTMC-reID. The training set of Market1501 contains 12,936 images captured under 6 cameras, and the training set of DukeMTMC-reID contains 16,522 images captured under 8 cameras. The number of person IDs captured by different cameras as well as the number of images are different, and their specific divisions are shown in Table 13.

Table 13. Distribution of different camera data in the dataset.

Camera ID	Market1501			DukeMTMC-reID		
	IDs	Sample Number	Rate (%)	IDs	Sample Number	Rate (%)
0	652	2017	3.09	404	2809	6.95
1	541	1709	3.16	378	3009	7.96
2	694	2707	3.90	201	1088	5.41
3	241	920	3.82	165	1395	8.45
4	576	2338	4.06	218	1685	7.73
5	558	3245	5.82	348	3700	10.63
6	-	-	-	217	1330	6.13
7	-	-	-	265	1506	5.68

We conduct small sample learning, using samples from each camera in the dataset as small sample training data, and used ResNet-50 network for experiments. It is generally believed that the larger the amount of training data, the better the training effect of the model. However, when the data volume is small, the overfitting phenomenon of the model is more obvious, and the experimental results are shown in Table 14. It can be seen that when using only single camera annotation, the performance results of training and testing on the Market1501 and DukeMTMC-reID datasets are poor. When training with Camera 2 data from the Market1501 dataset, the mAP reaches the highest 26.3%; when using the DukeMTMC-reID dataset for Camera 5 data training, the mAP reaches the highest

of 20.7%. Comparing Table 13 and Table 14, it can be found that there is no significant correlation between model performance and data size at this time. Both the Market1501 and DukeMTMC-reID datasets exhibit a significant decrease in performance due to the large amount of data collected by a certain camera. The data volume of Camera 5 in the Market1501 dataset is higher than that of Camera 2 and Camera 5, while the mAP is lower than that of Camera 2. Even compared with Camera 4, the data volume is increased by more mAP, but there is no significant improvement. The data volume of Camera 0, Camera 1 and Camera 3 is different, but the mAP is about 17%. The same is true for the DukeMTMC-reID dataset. Camera 2 has significantly less data than other cameras, and its performance is second only to Camera 5; Camera 1 has a large number of samples, and its mAP is only better than Camera 3 and Camera 4. The comparison results show that the diversity of data greatly affects the performance of the model. For small sample data, the more the sample coverage features, the more conducive to improve the generalization ability of the model.

Table 14. Single camera small sample learning experiment results.

Camera ID	Market1501				DukeMTMC-reID			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
0	17.7	43.2	61.4	69.0	18.1	37.7	50.9	57.3
1	17.4	41.7	60.2	67.9	16.4	32.9	47.8	53.9
2	26.3	54.2	72.7	78.9	19.6	40.5	54.7	61.2
3	16.8	40.3	59.6	67.5	14.2	31.4	45.3	51.0
4	21.1	47.7	65.7	73.1	15.4	32.4	47.2	52.5
5	22.7	49.3	68.1	75.2	20.7	40.4	55.2	57.7
6	-	-	-	-	17.4	35.8	51.9	57.7
7	-	-	-	-	17.6	36.4	51.1	57.3

Based on the single-camera labeling experiments, the non-target domain data are further added for testing. "M (single) + D" means that the single-labeled camera data of Market1501 is used as a small sample, and the complete DukeMTMC-reID is introduced for joint training and tested on Market1501; "D (single) + M" means that the single-labeled camera data of DukeMTMC-reID is used as a small sample, and the complete DukeMTMC-reID is introduced for joint training and tested on Market1501, respectively. DukeMTMC-reID's single-labeled camera data as small samples introduced the complete Market1501 for joint training, and tested on DukeMTMC-reID, and the results are shown in Table 15.

Table 15. Single-camera annotation with the addition of non-target domain data experimental results.

Camera ID	M (Single) + D (%)				D (Single) + M (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
0	25.4	54.0	70.9	77.4	24.2	42.2	58.3	64.4
1	24.2	52.1	68.1	74.8	21.0	36.4	52.2	58.9
2	29.2	58.8	74.3	80.2	23.5	40.4	56.8	64.2
3	25.0	53.0	70.3	76.8	20.3	36.4	51.8	59.0
4	27.0	54.3	72.5	78.5	21.4	37.8	53.1	59.7
5	27.7	55.8	73.6	80.2	23.8	42.4	57.0	63.4
6	-	-	-	-	22.5	39.7	54.8	60.6
7	-	-	-	-	22.6	39.4	55.1	61.6

As can be seen from Table 15, for both the Market1501 and DukeMTMC-reID datasets, any single camera annotation by adding non-target domain data is effective in improving the Re-ID performance of the model. In most cases, the better performance of the camera when only using a single camera annotation data, the improvement is not obvious after adding non-target domain data; on the contrary, the camera with poor performance when using only a single camera annotation data has a relatively more obvious improvement

after adding non-target domain data. It is fully demonstrated that different camera data have different coverage of sample diversity, and additional non-target domains can be supplemented to a certain extent. However, due to the influence of inter-domain differences, this auxiliary effect is not obvious. Even if a large amount of non-target domain data is added, it is still difficult for the model to learn sufficient information and knowledge from it.

In order to make more effective use of non-target domain data, the ResNet-50 network is now replaced with the model basic framework without changing other experimental details and parameter configuration. The effectiveness of the model basic framework (Figure 1) in this joint training scenario is verified by only changing the model architecture. The results are shown in Table 16. From the experimental data in the above tables, it is clear that the performance of Re-ID is significantly improved in various "M (single) + D" and "D (single) + M" scenarios. It indicates that the basic framework can fully utilize non-target domain data to assist in small sample training, and the model can extract pedestrian invariance features from additional data and combine small samples to improve performance. In the "M (single) + D" scenario, compared to adding non-target domain data, the mAP improvement of single camera annotation is 7.7%, 6.8%, 2.9%, 8.2%, 5.9%, and 5.0%, respectively. The mAP improvement is more significant and balanced by replacing the model architecture, which is 13.7%, 12.3%, 11.2%, 11.1%, 12.9%, and 11.8%, respectively; In the "D (single) + M" scenario, compared to adding non-target domain data, the mAP improvement of single camera annotation is 6.1%, 4.6%, 3.9%, 6.1%, 6.0%, 3.1%, 5.1%, and 4.0%, respectively. The mAP improvement is more significant and equally balanced and stable by replacing the model architecture, which is 12.2%, 12.7%, 10.0%, 9.9%, 10.9%, 11.6%, 10.9%, and 10.1%, respectively.

Table 16. Results of the joint training experiment using the base framework.

Camera ID	M (Single) + D (%)				D (Single) + M (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
0	39.1	67.4	83.0	87.6	36.4	55.3	71.6	76.8
1	36.5	52.1	80.1	84.7	33.7	53.2	67.3	72.5
2	40.4	69.3	82.7	87.1	33.5	52.9	68.1	72.9
3	36.1	66.1	81.3	86.2	30.2	48.9	63.6	69.5
4	39.9	68.0	83.2	87.6	32.3	51.7	66.1	71.1
5	39.5	68.6	83.2	87.9	35.4	56.0	68.9	74.0
6	-	-	-	-	33.4	53.1	67.6	73.2
7	-	-	-	-	32.7	50.0	66.3	72.5

In order to further improve the model performance, make full use of small sample data, and solve the problem of insufficient sample resolution learning, the model base framework is now replaced with the joint training framework proposed in this paper (Figure 4), while the same experimental details as well as parameter configurations are used for the experiments.The results are shown in Table 17.

Table 17. Results of the joint training experiment using the joint training framework.

Camera ID	M (Single) + D (%)				D (Single) + M (%)			
	mAP	Rank-1	Rank-5	Rank-10	mAP	Rank-1	Rank-5	Rank-10
0	48.2	75.4	86.8	90.3	40.3	61.8	74.6	79.4
1	43.6	71.2	83.9	88.5	42.5	62.7	73.3	77.2
2	54.9	79.0	89.4	82.2	41.4	62.6	75.4	79.4
3	43.0	71.5	85.2	89.3	35.7	56.8	70.2	75.0
4	48.5	75.4	87.3	90.7	40.7	61.1	72.9	77.9
5	47.9	75.9	88.4	92.3	43.5	64.1	76.1	80.3
6	-	-	-	-	41.7	63.1	74.1	77.7
7	-	-	-	-	41.9	62.8	75.0	79.3

Comparing Table 16 with Table 17, it is obvious that the experimental results of the joint training framework further increase substantially compared with the base framework. In the "M (single) + D" scenario, the maximum improvement of mAP is 14.5%, and the maximum improvement of Rank-1 is 9.7%; the minimum improvement of mAP is 6.9%, and the minimum improvement of Rank-1 is 5.4%. In the "D (single) + M" scenario, the maximum improvement of Re-ID evaluation index mAP is 12.7%, and the maximum improvement of Rank-1 is 9.2%; the minimum improvement of mAP is 9.9%, and the minimum improvement of Rank-1 is 3.9%. It fully illustrates the correctness of the framework design idea, and the joint training framework effectively solves the joint training problem of small sample data and non-target domain data. Both the learning bias caused by insufficient samples is solved by means of fusion features, and the inter-domain variation problem caused by additional samples introduced into the training is solved by RCFA modules.

In particular, the visual outputs for Market1501 and DukeMTMC-reID are shown in Figures 7 and 8. The first picture on the left side is the person to be matched, and the right side is the matching result from other cameras. It can be seen that the test results under our joint training framework can basically perform cross-domain matching under small samples.

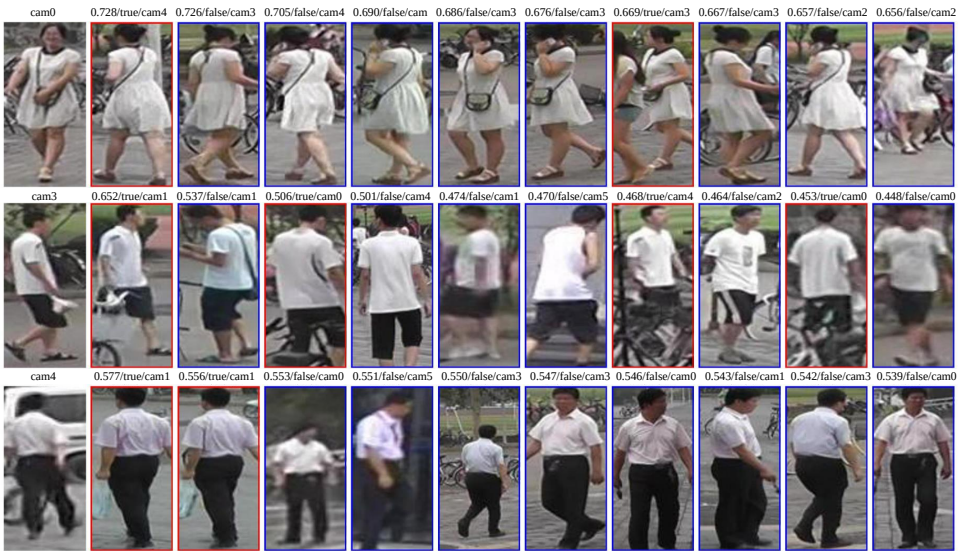

Figure 7. The matching results of Market1501.

Figure 8. The matching results of DukeMTMC-reID.

5. Conclusions

In this paper, we propose a Re-ID method based on RCFA, the core of which is an inserted residual compensation and fused attention module. This module can effectively improve the robustness of the model and overcome the perturbation caused by different data distribution to solve the Re-ID problem of cross-domain migration. In the cross-domain scenario from Market1501 to DukeMTMC-reID, the mAP metric improves by 6.3% and the Rank-1 metric improves by 10.1% compared to the baseline. The improvement of each component over the existing methods is also verified by comparing with existing attention mechanisms and style normalization methods through extensive experiments. In addition, a joint training framework combining cross-domain migration learning and small-sample learning is proposed, which can train both small-sample data and different domain data to effectively reduce the data collection and computational cost of realistic scenarios, while ensuring that the algorithm models can be adequately trained. Finally, the feasibility of this joint training framework is demonstrated through experiments in different scenarios. For example, 2017 images from camera 0 of the Market1501 dataset are selected as small samples and DukeMTMC-reID dataset are introduced for joint training, and tested on the Market1501 test set. The joint training framework improves mAP by 22.8% and Rank-1 by 21.4% compared to the ResNet-50 model alone. Our method assumes that the target domain class and the source domain class overlap to a large extent. In the actual scene, the target domain is likely to have rich source domain novel classes. How to further use these novel target domain classes is the problem we will explore next.

Author Contributions: Conceptualization, H.L.; methodology, K.Y.; software, Y.W.; validation, L.Z.and W.W.; formal analysis, H.L. and Y.L.; investigation, H.L.; resources, W.W.; data curation, W.W.;writing—original draft preparation, H.L.; writing—review and editing, Y.W.; visualization, L.Z.;supervision, G.Y. and Y.L.; project administration, G.Y.; funding acquisition, K.Y. and Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Natural Science Foundation of Xinjiang Uygur Autonomous Region (No. 2022D01B187 and No. 2022D01B05) and Shenzhen Science and Technology Program (No. JSGG20220301090405009).

Data Availability Statement: The figures and tables used to support the findings of this study are included in the article.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person re-identification: Past, present and future. *arXiv* **2016**, arXiv:1610.02984.
2. Gao, J.; Yang, X.; Zhang, T.; Xu, C. Robust visual tracking method via deep learning. *Chin. J. Comput.* **2016**, *39*, 1419–1434.
3. Zheng, Z.; Zheng, L.; Yang, Y. A discriminatively learned cnn embedding for person reidentification. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2017**, *14*, 1–20. [CrossRef]
4. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
5. Zheng, L.; Zhang, H.; Sun, S.; Chandraker, M.; Yang, Y.; Tian, Q. Person re-identification in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1367–1376.
6. Inoue, N.; Furuta, R.; Yamasaki, T.; Aizawa, K. Cross-domain weakly-supervised object detection through progressive domain adaptation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5001–5009.
7. Hou, L.; Zhang, Y.; Fu, K.; Li, J. Informative and consistent correspondence mining for cross-domain weakly supervised object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9929–9938.
8. Tang, Z.; Sun, Y.; Liu, S.; Yang, Y. DETR with Additional Global Aggregation for Cross-domain Weakly Supervised Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 11422–11432.
9. Xu, Y.; Sun, Y.; Yang, Z.; Miao, J.; Yang, Y. H2fa r-cnn: Holistic and hierarchical feature alignment for cross-domain weakly supervised object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14329–14339.
10. Gatys, L.A.; Ecker, A.S.; Bethge, M. A neural algorithm of artistic style. *arXiv* **2015**, arXiv:1508.06576.

11. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.
12. Zhu, J.Y.; Park, T.; Isola, P.; Efros, A.A. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2223–2232.
13. Chen, Y.; Zhu, X.; Gong, S. Instance-guided context rendering for cross-domain person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 232–242.
14. Chong, Y.; Peng, C.; Zhang, J.; Pan, S. Style transfer for unsupervised domain-adaptive person re-identification. *Neurocomputing* **2021**, *422*, 314–321. [CrossRef]
15. Zhu, Y.; Deng, C.; Cao, H.; Wang, H. Object and background disentanglement for unsupervised cross-domain person re-identification. *Neurocomputing* **2020**, *403*, 88–97. [CrossRef]
16. Song, J.; Yang, Y.; Song, Y.Z.; Xiang, T.; Hospedales, T.M. Generalizable person re-identification by domain-invariant mapping network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 719–728.
17. Zhou, Z.H. A brief introduction to weakly supervised learning. *Natl. Sci. Rev.* **2018**, *5*, 44–53. [CrossRef]
18. Hernández-González, J.; Inza, I.; Lozano, J.A. Weak supervision and other non-standard classification problems: A taxonomy. *Pattern Recognit. Lett.* **2016**, *69*, 49–55. [CrossRef]
19. Agrawala, A. Learning with a probabilistic teacher. *IEEE Trans. Inf. Theory* **1970**, *16*, 373–379. [CrossRef]
20. Blum, A.; Mitchell, T. Combining labeled and unlabeled data with co-training. In Proceedings of the Eleventh Annual Conference on Computational Learning Theory, Madisson, WI, USA, 24–26 July 1998; pp. 92–100.
21. Figueira, D.; Bazzani, L.; Minh, H.Q.; Cristani, M.; Bernardino, A.; Murino, V. Semi-supervised multi-feature learning for person re-identification. In Proceedings of the 2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance, Krakow, Poland, 27–30 August 2013; pp. 111–116.
22. Li, J.; Ma, A.J.; Yuen, P.C. Semi-supervised region metric learning for person re-identification. *Int. J. Comput. Vis.* **2018**, *126*, 855–874. [CrossRef]
23. Kodirov, E.; Xiang, T.; Fu, Z.; Gong, S. Person Re-Identification by Unsupervised l1 Graph Learning. In *Computer Vision—ECCV 2016: Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 178–195.
24. Liu, Z.; Wang, D.; Lu, H. Stepwise metric promotion for unsupervised video person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2429–2438.
25. Zhao, R.; Ouyang, W.; Wang, X. Unsupervised salience learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3586–3593.
26. Ye, M.; Li, J.; Ma, A.J.; Zheng, L.; Yuen, P.C. Dynamic graph co-matching for unsupervised video-based person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2976–2990. [CrossRef] [PubMed]
27. Wang, X.; Panda, R.; Liu, M.; Wang, Y.; Roy-Chowdhury, A.K. Exploiting global camera network constraints for unsupervised video person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *31*, 4020–4030. [CrossRef]
28. Fan, H.; Zheng, L.; Yan, C.; Yang, Y. Unsupervised person re-identification: Clustering and fine-tuning. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2018**, *14*, 1–18. [CrossRef]
29. Zeng, K.; Ning, M.; Wang, Y.; Guo, Y. Hierarchical clustering with hard-batch triplet loss for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 13657–13665.
30. Li, M.; Zhu, X.; Gong, S. Unsupervised person re-identification by deep learning tracklet association. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 737–753.
31. Xuan, S.; Zhang, S. Intra-inter camera similarity for unsupervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 11926–11935.
32. Fan, Q.; Zhuo, W.; Tang, C.K.; Tai, Y.W. Few-shot object detection with attention-RPN and multi-relation detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 4013–4022.
33. Chen, H.; Wang, Y.; Wang, G.; Qiao, Y. Lstd: A low-shot transfer detector for object detection. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 32.
34. Zhao, A.; Ding, M.; Lu, Z.; Xiang, T.; Niu, Y.; Guan, J.; Wen, J.R. Domain-adaptive few-shot learning. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021; pp. 1390–1399.
35. Nakamura, A.; Harada, T. Revisiting fine-tuning for few-shot learning. *arXiv* **2019**, arXiv:1910.00216.
36. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 770–778.
37. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
38. Luo, H.; Gu, Y.; Liao, X.; Lai, S.; Jiang, W. Bag of tricks and a strong baseline for deep person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.

39. 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]
40. 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, 7–13 December 2015; pp. 1116–1124.
41. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
42. 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, 21–26 July 2017; pp. 1318–1327.
43. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
44. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
45. 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, 18–23 June 2018; pp. 7132–7141.
46. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.