Article

An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification

Wei Huang ^{1,2} , Yongying Li ³, Kunlin Zhang ^{1,2} , Xiaoyu Hou ^{1,2}, Jihui Xu ^{1,2} , Ruidan Su ^{1,*}
and Huaiyu Xu ^{1,2,*}

- ¹ Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 201210, China; huangw@sari.ac.cn (W.H.); zhangkl@sari.ac.cn (K.Z.); houxiaoyu@sari.ac.cn (X.H.); xujh@sari.ac.cn (J.X.)
² University of Chinese Academy of Sciences, Beijing 100049, China
³ School of Mechanical Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 182371459@st.usst.edu.cn
* Correspondence: surd@sari.ac.cn (R.S.); xuhuaiyu@sari.ac.cn (H.X.)

Abstract: The multi-scale lightweight network and attention mechanism recently attracted attention in person re-identification (ReID) as it is capable of improving the model's ability to process information with low computational cost. However, state-of-the-art methods mostly concentrate on the spatial attention and big block channel attention model with high computational complexity while rarely investigate the inside block attention with the lightweight network, which cannot meet the requirements of high efficiency and low latency in the actual ReID system. In this paper, a novel lightweight person ReID model is designed firstly, called Multi-Scale Focusing Attention Network (MSFANet), to capture robust and elaborate multi-scale ReID features, which have fewer float-computing and higher performance. MSFANet is achieved by designing a multi-branch depth-wise separable convolution module, combining with an inside block attention module, to extract and fuse multi-scale features independently. In addition, we design a multi-stage backbone with the '1-2-3' form, which can significantly reduce computational cost. Furthermore, the MSFANet is exceptionally lightweight and can be embedded in a ReID framework flexibly. Secondly, an efficient loss function combining softmax loss and TriHard loss, based on the proposed optimal data augmentation method, is designed for faster convergence and better model generalization ability. Finally, the experimental results on two big ReID datasets (Market1501 and DukeMTMC) and two small ReID datasets (VIPeR, GRID) show that the proposed MSFANet achieves the best mAP performance and the lowest computational complexity compared with state-of-the-art methods, which are increasing by 2.3% and decreasing by 18.2%, respectively.

Keywords: person ReID; multi-scale feature learning; attention mechanism; lightweight network; softmax and TriHard loss

Citation: Huang, W.; Li, Y.; Zhang, K.; Hou, X.; Xu, J.; Su, R.; Xu, H. An Efficient Multi-Scale Focusing Attention Network for Person Re-Identification. *Appl. Sci.* **2021**, *11*, 2010. <https://doi.org/10.3390/app11052010>

Academic Editor: Manuel Armada

Received: 5 January 2021

Accepted: 12 February 2021

Published: 25 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Person re-identification (ReID) is a computer vision task that gives image information [1] or video sequence, and even text feature description [2] about a query of a person of interest, to determine where this person has appeared in another place at a distinct time through different cameras. ReID is developed to make up for the visual limitations of the fixed camera, which can combine with pedestrian attribute recognition, pedestrian detection, and pedestrian tracking technologies, widely applied to intelligent video surveillance, intelligent security, and edge intelligence, etc. Currently, challenges of person re-identification mainly derive from the presence of different viewpoints [3], varying low-image resolutions [4], changing illumination [5], and unconstrained poses [6], etc. These stringent requirements constitute an insuperable challenge for ReID methods based on manual features. Deep learning is a promising approach to complete the person ReID task by employing an end to end neural network. ReID is more practical than ever before under

the booming development of computer vision, progressively transition from a closed-world setting centered around theoretical research to an open-world stage aimed at actual landing applications [7]. Specifically, the development of person ReID is moving from the same domain to the cross-domain [8], bounding box image to raw image [9], the gallery contains the query to open-set [10], etc. Besides, person ReID based on deep learning usually faces ethical and privacy issues. Deep learning for person ReID also evolves from heavyweight to lightweight to meet the computing core's performance. Ref. [11] use Light-weight deep learning model and IoT collaboration to pre-process surveillance video. Mittal et al. [12] have introduced many optimized and efficient deep learning models to the NVIDIA Jetson platform. In addition, ReID models are increasingly deployed in the mobile environment; however, the computing limitation of supporting hardware is another challenge of edge deployment. Therefore, designing an efficient and lightweight ReID model is an essential aspect of facilitating the application of ReID.

The designs of ReID network mostly refer to feature learning ideas from image classification networks, such as ResNet [13], MoblieNet [14], ShuffleNet [15], BraidNet [16], Inception [17], and SeNet [18], etc. For mobile deployment, deep neural network model compression acceleration and computing power cost are required because of the huge model parameters. Recently, some methods based on deep learning multi-layer neural network, combined with lightweight network, attention mechanism or multi-scale feature fusion module, have been widely applied in ReID to learn distinctive features [19–27]. However, most of these methods have not dealt with the integration of global features and local features based on above strategies in ReID task properly, resulting in learning redundant ReID features and higher computation complexity. For the excellent lightweight OSNet [27], it relied on depthwise separable convolution and a novel unified aggregation gate to extract omni-scale features, which achieves state-of-the-art performance on six open person ReID datasets. However, it is limited to learn big block attention features and neglects the inside convolution layer's ones. Moreover, the edge deployment often relies on third parties to accelerate the optimization algorithm to improve the inference speed. Presently, the ReID model is mainly designed around high efficiency and lightweight to meet edge deployment requirements.

To address the aforementioned problems and challenges of ReID methods, a novel ReID model, called the Multi-Scale Focusing Attention network (MSFANet), is proposed in this paper, which has three stages of different module sizes. Every stage is composed of a different number of multi-scale attention modules, which can be applied to extract the ReID features of multi-scale receptive fields efficiently. Similar to the lite3*3 module in OSNet [27], we use depthwise separable convolution [14,28] to reduce model computations and parameters. However, OSNet has only an adaptive aggregation after the multi-scale receptive field module blocks, which do not extract attention features inside each block. There is evidence that elaborative attentions play a crucial role in regulating ReID feature [19–23,27]. Therefore, it is hard to obtain more in-depth and robust ReID features in each OSNet's stage. We refer to the SENet [18] structure and design a new lightweight focusing-attention (FA) block, using both the depthwise separable convolution and focusing channel attention module inside each FA block to extract more robust ReID features. Due to the sizes of a person in different cameras are usually quite discriminative, we use more extensive cross-scale ReID features in each MSFA module to reduce the number of branches. Later experimental results have shown that our model design dramatically reduces the model computations and improves mAP and Rank-1 metrics. To alleviate the problems of target occlusion, deformation, blur, etc., the optimal data augmentation method, namely the combination of flipping, cropping, erasing, patching and autoaugment [29], is obtained through a large number of experiments in this paper, which makes MSFANet more robust. Besides, we combine the softmax loss function for classification tasks and the TriHard [30] loss function for hard sample discrimination to build a new fusion loss function. Experiments show that the loss function helps the training process with faster convergence on MSFANet and improves the generalization ability of this model.

Our contributions can be summarized as follows:

- The Multi-Scale Focusing Attention Network (MSFANet), containing receptive fields of different sizes in each stage as focusing attention (FA) block, is proposed to capture multiple scales and highly focusing attention ReID features. It is a novel efficient network module that achieves lower computational complexity and can be flexibly embedded in a person ReID framework.
- We design a new fusion loss function to address hard sample discrimination problems, which is a combination of softmax loss and hard triplet loss. It achieves faster convergence and better generalization ability of the proposed MSFANet.
- Experiments on the Market1501 [1], DukeMTMC-ReID [31], and two small ReID datasets (VIPeR [32], GRID [33]) show the superiority of the proposed MSFANet over a wide range of state-of-the-art ReID models in terms of model accuracy and computational complexity.

2. Related Work

2.1. Person ReID and Attention Mechanism

With the rise of deep learning, there has been an increasing amount of literature on person re-identification, such as deep metric learning [30,34] feature representation learning [27,35], attribute learning [36], and re-ranking techniques [37]. Taken together, these studies presented different types of ReID methods to address a wide variety of ReID problems or challenges. Furthermore, attention methods have achieved great success in ReID [19–23]. In this paper, we aim to introduce an improved person re-identification method via the attention mechanism to enhance the performance of ReID.

The basic idea of the attention mechanism in computer vision is to allow the deep learning model to ignore irrelevant information and focus on critical details. ASTPN [20] (A joint Spatial and Temporal Attention Pooling Network) is based on the attention mechanism to learn ReID features in video sequences better. The video sequence-based ReID task has the problem of varying lengths of video sequences. Chen et al. [21] split a long video sequence into multiple short segments and merged the first few segments after sorting for model learning. Based on the clip-level spatial and temporal attention (STA) [22] learning features of spatial and temporal dimensions. The self-attention mechanism is also used to learn the relationship between long and short video sequences [21]. ABD-Net [19] is designed to combine attention modules and diversity as complementary mechanisms to improve ReID performance jointly. Ref. [23] proposed an effective model to determine the importance of each spatial-temporal feature from a global perspective. However, these spatial-temporal attention methods are coarse and unstable, which cannot deal with the complex relationships among parts. Thus, the Multi-Scale Focusing Attention Network (MSFANet) is proposed to capture more complex and detailed information for ReID training task.

2.2. Lightweight Network Designs

To meet edge deployment's performance requirements, lightweight convolutional neural network designs have recently attracted special attention. Inceptionv3 [17] and Xception [38] reduce the network's parameters via depthwise separable convolution. MobileNet [14] leverage depthwise separable convolution and inverted residuals, respectively. ShuffleNet [15] used pointwise group convolution and channel shuffle to reduce model computational cost. AdderNet [39] proposed an adder network to replace massive multiplications in the deep convolutional neural network. OSNet [27] has achieved great success in lightweight network design for ReID task, which is analogous to MobileNet [14] and Xception [38] by employing depthwise separable convolution. Therefore, in this paper, our proposed MSFANet is also adopting this strategy to reduce model computational cost, and combined with attention mechanism, while outperforming OSNet.

2.3. Multi-Scale Feature Fusion Learning

The way of capturing targets' features in a convolutional neural network usually adopts layer-by-layer extraction in the abstract. Over the past decade, most research in computer vision tasks, such as object detection, image segmentation, person re-identification (ReID), and so on, emphasized the effectiveness of multi-scale feature learning [27,40–42], which is relevant to the receptive field. FPN [35] solves multi-scale problems by fusing multi-level features. For object detection tasks, refs. [42–44] carried out many investigations into the multi-scale target detection-based FPN to achieve higher accuracy. Based on previous studies of FPN, ref. [45] proposed an optimization strategy of bottom-up path augmentation to improve the use of low-level information and accelerate the dissemination efficiency of low-level details. NAS-FPN [40] employed neural architecture search with reinforcement learning to automatically generate a feature pyramid structure. Tan et al. [41,42] investigated the Bidirectional Feature Pyramid Network (BiFPN) to achieve multi-scale target classification and detection, which introduces learnable weights to the input to learn the importance of different input features, similar to the attention mechanism, and the weight is able to apply to all features or channels. Ref. [46] proposed an attention mechanism to learn the soft weight of the multi-scale features at each pixel location. Ref. [47] proposed a new semantic attention segmentation model that aggregates multi-scale and context information to re-predict the finer. Ref. [48] presented hierarchical multi-scale attention for semantic segmentation model, which reduces memory consumption and improves network training speed and prediction accuracy. For person ReID tasks, refs. [26,49] studied a multi-layer structure to extract features that perform best under different convolution depths. OSNet [27] carried out quite different methods to obtain the features of multi-scale receptive fields, which adaptively aggregated four branches with varying numbers of lightweight blocks to get multi-scale features but aggregated features of different network layers. It dramatically reduces network parameters and computations. Overall, these studies highlight the importance and necessity of multi-scale feature learning. This paper's proposed MSFANet draws on the OSNet's architecture, which employs fewer branches and larger receptive field features to further improve ReID's performance.

3. Methods

3.1. Pipeline

In this part, we introduce the framework of the Multi-Scale Focusing Attention Network (MSFANet) for ReID tasks, shown in Figure 1, which contains three main parts as data augmentation, backbone, and two inference heads. Firstly, we will introduce the best data augmentation methods for data preprocessing (Section 3.2). Secondly, the proposed MSFANet's backbone and head will be described in detail (Section 3.3). Finally, a fusion loss function is designed to train our model properly (Section 3.4).

Figure 1. The architecture of Multi-Scale Focusing Attention Network (MSFANet), containing four steps: (1) Input of training images; (2) Data augmentation on training images; (3) The backbone of this model for extracting images' feature, which is composed of three stages with 1, 2, and 3 MSFA Blocks; (4) Heads of MSFANet for training with softmax and triplet loss function, respectively.

3.2. Data Augmentation

Data augmentation is used to introduce prior knowledge for visual invariance or semantic invariance in computer vision by carrying out image transformations on color, geometric, morphologic, etc., which can play a regularization role in the model training and has become an essential step to improve the model's performance.

For ReID task, it is crucial to leverage data augmentation to help the model learn more discriminative features. Considering the person ReID task in the actual scene, there are many factors such as occlusion, viewpoint change, and background change, etc. Combining a mass of homologous experiments in this paper, we found that data augmentation is extremely beneficial to improve the model's generalization capacity by performing random flipping, random cropping, random erasing, and random patching on varieties of ReID datasets. Specifically, as shown in Figure 2, the random flipping's operation is randomly flipping the image in the horizontal direction. In addition, random cropping is a method of cutting edges randomly, helping to focus on the area of interest, which can be seen in column 2. Besides, the random erasing and patching are both used to imitate target occlusion in the actual scene. Furthermore, we also add autoaugment [29], applying reinforcement learning, to learn data augmentation rules from datasets, which can be very convenient to transfer to new datasets. The following experimental results verify that together these data augmentations can achieve comprehensive performance improvement based on MFSFANet.

Figure 2. Examples of the results of different data augmentation methods on MSFANet. These methods contain random flipping, random cropping, random erasing, random patching, and autoaugment, shown in columns 2 through 5, respectively.

3.3. Multi-Scale Focusing Attention Network

In this section, the architecture of MFSFANet is depicted in detail, which specializes in learning multi-scale feature representation for a person ReID task. We mainly introduce the backbone of MFSFANet, following two heads as full connected (FC) layer for classifica-

tion and global average pooling layer for capturing person's ReID features, respectively. The backbone consists of a standard 7×7 convolution and average pooling layer, connecting three multi-scale focusing attention with different quantities and sizes. The benefits of this design are that it cannot only extract higher robustness multi-scale person ReID features, but also has much lower computational costs, compared with OSNet [27].

Focusing Attention (FA) Block. Attention is a more efficient way of capturing more robust multi-scale features in computer vision tasks, which is often used in transition block or aggregation block after multiple branches network rather than focusing on the inside of a feature extraction module in each branch. The FA block, shown in Figure 3a, is a lightweight feature extraction block with attention in the proposed MSFANet. FA block contains depthwise separable 1×1 convolutions and 3×3 convolutions in the front, then connects batch normalization and Leaky_ReLU layers, and binds a Squeeze-and-Excitation (SE) block [18] at the back, shown in Figure 3b. A significant advantage of the FA block is that it can extract more elaborative ReID features inside the block, which can be embedded into a subnet of deep neural network expediently. Furthermore, it is a relatively lightweight design that can reduce the computational cost in training and inferencing. SE block is referred to SENet [18], while we replace Relu with Leaky_ReLU in the activation layer. In our experiments, this small change can provide a better convergence during the training process. The SE block principles are leveraging two fully connected (FC) layers to reduce and increase the channel's dimension, respectively, obtaining the attention weight w_c of the input x , and then employing the channel-wise multiplication between the w_c and the input x . Thus, the FA block provides a means of gaining robust attention features based on depthwise separable convolution, which will primarily be employed to the MSFANet.

Figure 3. An illustration of Focusing Attention (FA) block as (a) and Squeeze-and-Excitation (SE) block as (b). FA block consists of depthwise separable convolution and SE block, the latter's architecture is shown in (b).

Multi-Scale Focusing Attention (MSFA) Block. In order to achieve multi-scale feature learning, we design a new multi-branch structure with residual connection based on FA blocks, called Multi-Scale Focusing Attention (MSFA) block, shown in Figure 4. The block has three branches, each of which has 1, 3, 5 FA blocks that corresponding receptive fields are 3, 7, 11, respectively. This design can effectively extract the features of bigger cross-scale targets. Besides, we refer to the Aggregation Gate (AG) module of OSNet [27] to fuse the features of different branches. Thus, our proposed MSFA Block is capable of

capturing much more discriminative and robust features for ReID. The input x_l is from the last layer's output, $x_l \in R^{N \times C \times H \times W}$. The MSFA block aims to learn a residual feature with a mapping function $\mathcal{F}$:

$$x_{l+1} = x_l + \mathcal{F}(x_l, W_l). \quad (1)$$

where l indicates at which stage and $\mathcal{F}$ denotes one stage's aggregation features of multi-branch FA layer that learn multi-scale fusion features, and $\mathcal{F}$ can be obtained by Equation (2), is

$$\mathcal{F} = \sum_{t=1}^T FA_t(x_l^t, W_l^t). \quad (2)$$

where T is the number of MSFA block's branches, x_l^t and W_l^t represent the input and weights of the t th branch in the l th stage, respectively. FA_t denotes the focusing attention features of the t th branch, FA'_{t+1} is the output of the $(2 * t - 1)$ th FA block in the $(t + 1)$ th branch. This operation can combine two different branch features before the final aggregation layer, taking the sum and average of the two branches. As shown in Equation (3), G is referred to as the Aggregation Gate (AG) of OSNet [27], which is a learnable mini-network to achieve the dynamic scale-fusion and $\odot$ represents the Hadamard product.

$$FA_t = \frac{1}{2}(FA_t + FA'_{t+1}) \odot G(x_1^t), \quad \text{s.t. } t \geq 1 \text{ and } t < T \quad (3)$$

Figure 4. An illustration of the MSFA block, which aggregates three branches with 1, 3, and 5 FA blocks, introduced in Figure 3, to extract multi-scale receptive field features.

The backbone of MSFANet is shown in Figure 1, mainly composed of three stages, including 1, 2, and 3 MSFA blocks. We set different output channels for three stages with 256, 384, and 512 in our experiments. The backbone form can also be set as '2-2-2', which means that it has two MSFA blocks in each stage. For backbone as the form of '1-2-3', we compare our MSFANet backbone with OSNet [27] and found that the computational complexity reduced by 18% while the mAP increased by 1.5%. The detailed results can be viewed in the experimental section.

3.4. Loss Function

Our proposed MSFANet is trained in an end-to-end manner. For person re-identification, it can be considered as a classification task. Various studies have assessed the influence of loss function for a ReID task. In this work, we adopt the widely used softmax loss and Triplet loss with batch hard mining to train MSFANet in opening ReID datasets.

Softmax loss. For a classification training task, softmax loss is very suitable, regarded as a cross-entropy loss using the softmax function. It is defined in the following Equation (4), where S_j is the j th value of the output vector S of the softmax layer, representing the probability that this sample belongs to the j th category. y_j represents a $1 * T$ vector,

and T denotes the number of categories, of which only one value is 1, and the other $T - 1$ values are all 0.

$$L_{softmax} = \sum_{j=1}^T y_j \log S_j. \quad (4)$$

TriHard Loss [30]. Triplet loss with batch hard mining is an improved version of triplet loss, which is able to improve the generalization ability of a ReID network. TriHard loss's mathematical form is shown in Equation (5), where P denotes the number of different person IDs randomly selected, and K represents the number of pictures randomly chosen for each person ID. The goal is that for each picture a in the batch to select a hardest positive sample and a hardest negative sample to form a triple with it.

$$L_{tri_h} = \frac{1}{P \times K} \sum_{a \in batch} (\max_{p \in A} d_{a,p} - \min_{n \in B} d_{a,n} + \alpha) \quad (5)$$

Fusion Loss. Softmax loss and TriHard loss both have their respective advantages in ReID model's training. However, there are some problems when we use softmax loss alone. As presented in the top part (1) and middle part (2) of Figure 5, we can observe that the softmax loss can distinguish positive and negative sample well to some extent, but is hard to deal with some hard samples. TriHard loss can achieve batch hard sample mining during model training. It can shorten the distance from hard positive samples while maximizing the distance of hard negative pairs [30]. In this work, we combine the two loss functions to improve the generalization of MSFANet, which can alleviate the hard sample differentiation problem shown in bottom part (3) of Figure 5. The fusion loss can be viewed as Equation (6), where α and β are hyperparameters that can adjust the above two loss functions' weight. In our abundant experiments in different ReID datasets, we found that MSFANet can achieve the best robustness and fastest convergence when α and β are set 1 and 0.5, respectively. Thus, our subsequent experiments are all based on this parameter setting.

$$L = \alpha L_{softmax} + \beta L_{tri_h} \quad (6)$$

Figure 5. An illustration of comparing different loss functions applying to training data. The top (1) shows the result of the softmax loss on easy samples, while the middle (2) and bottom (3) show the differential results of the softmax loss and fusion loss on hard samples.

4. Experiments

In this section, extensive experiments of the proposed MSFANet on varieties of ReID datasets are demonstrated. The experiment's hardware environment is characterized by Inter Core i7-6500k CPU 3.4 GHz with two GPUs of TITAN RTX 24 G memory, while the software environment is configured by Python3.7, Pytorch1.7.0, and Torchreid [50] framework.

4.1. Implementation Details

Datasets and Evaluation Metric. Our experiments were carried out on two big datasets (Market1501 [1], DukeMTMC [31]) and two small datasets (VIPeR [32], GRID [33]) to verify MSFANet's effectiveness. Table 1 shows the number of cameras, train data, query data, and gallery data for different datasets. The Cumulative Matching characteristic CMC (Rank-1) and mAP were used as the model's evaluation metrics in this experiment. Rank-n represents the probability that the top n images in the search results (with the highest confidence) have correct matching, while mAP measures the model's performance across all categories.

Table 1. The statistics of experimental datasets. (T: Train; Q: Query; G: Gallery.)

Dataset	Time	Camera	Resolution	IDs (T-Q-G)	Images (T-Q-G)
Market1501 [1]	2015	6	fixed	751-750-751	12936-3368-15913
DukeMTMC [31]	2017	8	fixed	702-702-1110	16522-2228-17661
VIPeR [32]	2007	2	fixed	316-316-316	632-632-632
GRID [33]	2009	8	vary	125-125-900	250-125-900

Training process. To make a completed comparison with existing models, two training visions of MSFANet are provided. The first one is starting from scratch on the corresponding dataset without pre-training. The second is using ImageNet dataset for pre-training. The batch size and weight decay are set as 128 and 5×10^{-4} , respectively. Random sampling and AMSGrad are chosen for optimizing. In addition, we set the maximum training epochs as 400. The initial learning rate is 0.065 and is decayed by 0.1 at 150, 225, 300, and 350 epochs. For the ImageNet pre-training vision, the initial learning rate was set as 0.0015 and a cosine learning rate with adjustment strategy was used. Besides, the first ten epochs only train the classification layer, then freezing the classification layer and fine-tune the other layers at a lower learning rate.

4.2. Results

Data Augmentation Ablation Analysis. It can be seen from the data in Table 2 that the data augmentation methods have a significant impact on the performance of MSFANet. We test the model's mAP and Rank-1 with different data augmentation method combinations, including flipping, cropping, erasing, color jitter, rotation, and autoaugment, etc. We set flipping, cropping, erasing, and patching as a basic data augmentation combination, which can be combined with color jitter, affine, rotation, cutout, and autoaugment, respectively. As shown in Figure 6a, we compare these different combinations' mAP. We can find that the best performance can be achieved when adopting a novel combination of basic data augmentation with autoaugment. As shown in the last row of the following table, it can achieve 82% mAP and 92.5% Rank-1 that obviously surpass other data augmentation combinations. Therefore, in the subsequent experiments, this data augmentation combination is carried out for the proposed MSFANet training.

Results of loss function. To verify the proposed fusion loss function's effectiveness, illustrated in Equation (6), we first set α as a decimal between 0 and 1 and β as 1 to find the optimal coefficient α . Figure 6b compares the mAP based on fusion loss functions with different coefficients, and it can be observed that the mAP achieves the best when α is set to 0.5. Then, in order to further verify this fusion loss function, we train MSFANet with softmax loss function and fusion loss function on Market1501 and DukeMTMC datasets, respectively. As shown in Table 3, the mAP and Rank-1 increase 0.6% and 0.3% on Market1501 dataset while increase 2.2% and 0.9% on DukeMTMC datasets. These results thus demonstrate the effectiveness and necessity of our fusion loss function.

Table 2. Results of data augmentation methods on MSFANet. The mAP and Rank-1 of this model are calculated by combining different data augmentation methods.

MSFANet Data Augmentation on Market1501										mAP (%)	Rank-1 (%)
Horizontal flip	Crop	Erase	Patch	Color jitter	Affine	Rotation	Cutout	AutoAugment		81.3	91.0
Y	Y	Y	Y	Y						81.8	91.1
Y										82.2	92.4
Y	Y									83.1	92.5
Y	Y	Y	Y			Y				83.5	92.6
Y	Y	Y	Y		Y					83.9	92.8
	Y	Y	Y	Y						84.6	93.3
Y	Y	Y	Y							85.0	93.9
Y	Y	Y	Y	Y						85.1	93.9
Y	Y	Y	Y				Y			86.0	94.5
Y	Y	Y	Y					Y		86.8	94.6

(a) mAP-Epochs (Data Augmentation)

(b) mAP-Epochs (Loss Function)

Figure 6. mAP (%) under epochs based on different data augmentation combinations (a). mAP (%) under epochs based on fusion loss function with different coefficients (b), where $T\alpha$ denotes the result of the fusion loss function with weight α .

Table 3. Results of different loss functions on MSFANet. mAP and Rank-1 are calculated on Market1501 and DukeMTMC datasets.

Loss	Dataset	mAP	Rank-1
Softmax	Market1501	86.2	94.3
0.5*Triplet + Softmax	Market1501	86.8	94.6
Softmax	DukeMTMC	74.4	85.9
0.5*Triplet + Softmax	DukeMTMC	76.6	86.8

Results on big ReID datasets. As shown in Table 4, the experimental results of MSFANet on big ReID datasets are presented, including the model's complexities and inference abilities. The table compares the performance of MSFANet with other state-of-the-art methods, which contains top and bottom two parts, representing training from scratch or not, respectively.

- **Model complexity analysis.** The complexity of ReID model, affecting the training and computational cost, plays a vital role in the edge deployment. From the last column of Table 4, it can be observed that MSFANet outperforms most published methods about the model's complexity by a clear margin. The parameters and GFLOPs of MSFANet are calculated on resized input images by 256*128, and we can see that the GFLOPs is only 0.82 when the model's backbone is '1-2-3' form, decreased by 18.2% from OSNet [27]. This design verifies the effectiveness of the multi-scale focusing attention network for ReID by applying a depthwise separable convolution mechanism. However, as the backbone form is set as '2-2-2', the GFLOPs will be increased to 0.94, almost equal to OSNet's. Thus, MSFANet's backbone's structural design also plays an important role in the model's computational complexity. Compared to other models, MSFANet is a lightweight model, with significantly smaller parameters and floating-point computations than other heavyweight models (such as ResNet for the backbone). It performs better even compared with OSNet, regarded as a state-of-the-art lightweight ReID model by far.
- **Model Accuracy analysis.** It can be seen from the data in Table 4 that the Rank-1 and mAP of MSFANet and other state-of-the-art methods on Market1501 and DukeMTMC datasets are calculated. The top half of the table, training model from scratch, show that the mAP has increased by 1% and 1.3% when the backbone's structures are '1-2-3' and '2-2-2'. For the bottom half table, training with ImageNet pre-training, the mAP of MSFANet employing the two backbone structures is 86.8% and 86.9%, respectively, huge improvements over OSNet. The results obtained from the preliminary analysis of MSFANet's high efficiency are shown in Table 4. The designs of MSFANet contain multi-scale receptive fields, Omni-scale feature learning, and channel-level attention mechanism, which can capture much more robust and meticulous ReID features. MSFANet uses fewer branches and basic modules, integrating the attention module, which helps speed up the inference of ReID.

Overall, these results in Table 4 indicate that MSFANet achieves a better trade-off between model lightweight and model inference capabilities than other lightweight models (such as OSNet, ShuffleNet, and MobileNet, etc.). To some extent, it saves a lot of memory and computing resources on account of the high performance of MSFANet.

Results on small ReID datasets. GRID and VIPeR are two tiny ReID datasets. There are only a few hundred training images. Training a ReID model from scratch on these two datasets is a considerable challenge. For a fair comparison, we use transfer learning like other methods. MSFANet uses the pre-trained weights on the Market1501 to perform fine-tuning on these two datasets. As shown in Table 5, compared with OSNet, we can observe that the Rank-1 and mAP of MSFANet with our new data augmentation method increase 10.1% and 10% on VIPeR datasets, respectively. On GRID dataset, MSFANet's mAP is increasing by 0.8%. When our model adopts the same data enhancement method as OSNet, as shown in the penultimate row of Table 5, the mAP and Rank-1 indicators are also greatly improved. Thus, MSFANet also has better performance on small ReID datasets.

Table 4. Results (%) on big ReID datasets. The top half of the table shows the performance of MSFANet without ImageNet pre-training, compared with other state-of-the-art methods, while the bottom half are the results based on ImageNet pre-training. (‡: Utilizing attention mechanisms. †: Train from scratch).

Method	Publication	Backbone	Market1501		DukeMTMC		Complexity	
			Rank-1	mAP	Rank-1	mAP	Param (10 ⁶)	GFLOPs
Without ImageNet Pre-training								
† ShuffleNet [15]	CVPR'18	ShuffleNet	84.8	65	71.6	49.9	5.34	0.38
† MobileNetV2 [14]	CVPR'18	MobileNetV2	87	69.5	75.2	55.8	4.32	0.44
† BraidNet [16]	CVPR'18	BraidNet	83.7	69.5	76.4	59.5	-	-
‡ HAN [25]	CVPR'18	Inception	91.2	75.7	80.5	63.8	4.52	0.53
† PCB [24]	ECCV'18	ResNet	92.4	77.3	81.9	65.3	23.53	2.74
† OSNet [27]	ICCV'19	OSNet	93.6	81	84.7	68.6	2.23	0.98
‡ MSFANet + '1-2-3' (ours)	-	MSFANet	92.5	82	82.6	69.4	2.52	0.82
‡ MSFANet + '2-2-2' (ours)	-	MSFANet	92.9	82.3	82.8	69.5	2.25	0.94
With ImageNet Pre-training								
DaRe [4]	CVPR'18	DenseNet	89	76	80.2	64.5	6.95	1.85
MLFN [26]	CVPR'18	ResNeXt	90	74.3	81	62.8	22.98	2.76
‡ Bilinear [51]	ECCV'18	Inception	91.7	79.6	84.4	69.3	4.5	0.5
G2G [52]	CVPR'18	ResNet	92.7	82.5	80.7	66.4	23.5	2.7
DeepCRF [53]	CVPR'18	ResNet	93.5	81.6	84.9	69.5	23.5	2.7
IANet [54]	CVPR'19	ResNet	94.4	84.5	87.1	73.4	23.5	2.7
DGNet [55]	CVPR'19	ResNet	94.8	84.5	86.6	74.8	23.5	2.7
OSNet [27]	ICCV'19	OSNet	94.8	84.5	88.6	73.5	2.2	0.98
‡ MSFANet + '1-2-3' (ours)	-	MSFANet	94.3	86.8	86.8	76.6	2.52	0.82
‡ MSFANet + '2-2-2' (ours)	-	MSFANet	94.5	86.9	87.0	76.3	2.25	0.94

Table 5. Results (%) on small ReID datasets. Training OSNet and MSFANet from scratch on the two datasets. (‡: Utilizing attention mechanisms. †: Train from scratch.)

Comparison to State-of-the-Art Methods on Small ReID Datasets						
Method	Publication	Backbone	VIPeR		GRID	
			Rank-1	mAP	Rank-1	mAP
JLML [56]	CVPR'18	ResNet	50.2	-	37.5	-
HydraPlus-Net [57]	CVPR'18	Inception	56.6	-	-	-
OSNet [27]	ICCV'19	OSNet	41.1	54.5	38.2	40.5
‡ MSFANet - data_aug (ours)	-	MSFANet	48.9	60.3	28.9	40.8
‡ MSFANet + data_aug (ours)	-	MSFANet	52.2	64.5	30.4	41.3

Results on Feature learning of MSFANet. We conduct a set of feature map visualizations on the final output feature maps of MSFANet with attention, MSFANet without attention, OSNet, MobileNet v2, and ResNet50 with attention, as shown in Figure 7. We notice that MSFANet without attention has no apparent attention to region features, while MSFANet with attention can better focus on the regional features, as shown in the 1st and 2nd columns of the following figure. However, the attention regions can sometimes overly emphasize some local areas (e.g., clothes), implying the risk of overfitting person-irrelevant nuisances. In contrast, the attention of MSFANet can strike a better balance: it focuses on more of the local parts of the person's body while still being able to eliminate the person from backgrounds. Compared with other models, such as OSNet, MobileNet v2, and ResNet50, MSFANet's multi-scale attention has a more preminent ability in extracting discriminative ReID features from the row images. MobileNetV2 and ResNet50 are more about capturing global features and ignoring significant local features. The results show that MSFANet performs state-of-the-art.

Results on the inference of MSFANet. In the inference stage, given a query image, MSFANet can find the same ID in the gallery in a short time. The gallery can be a messy image collection or a series of video sequences. As shown in Figure 8, we test some samples on MSFANet. It can be observed that no matter for simple samples (1), (2), (3), (5), or for difficult samples (4), MSFANet can return the correct match (if the query in gallery).

Figure 7. Visualization of feature maps based on different feature learning methods. (i) MSFANet with attention; (ii) MSFANet without attention; (iii) OSNet: Omni-scale feature; (iv) MobileNet v2 and ResNet50: Global feature learning method. (1)–(6) show the feature visualization results of different test samples.

Figure 8. The inference results of MSFANet. For each query, the top ten searching results from the gallery are shown. The green box indicates the correct match, while the red box is the opposite. (1)–(5) show the query results for the different test samples.

5. Conclusions

In this paper, we first propose a novel and efficient ReID model, called Multi-Scale Focusing Attention Network (MSFANet), which is a lightweight and an attention-based net that is capable of learning multi-scale feature representations and achieving a better trade-off at the computational complexity and accuracy of the model. Then we design different data augmentation methods on MSFANet and obtain the final optimal combination. Next, developing a new fusion loss function by adding softmax loss and TriHard loss weighted. These improvements are conducive to promoting the development of ReID technology. In addition, extensive experiments have been conducted over four opening ReID datasets that validating the necessity and effectiveness of our method.

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

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data available on request due to restrictions e.g., privacy or ethical. The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy.

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

References

1. 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.
2. Li, S.; Xiao, T.; Li, H.; Yang, W.; Wang, X. Identity-aware textual-visual matching with latent co-attention. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1890–1899.
3. Karanam, S.; Li, Y.; Radke, R.J. Person re-identification with discriminatively trained viewpoint invariant dictionaries. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 4516–4524.
4. Wang, Y.; Wang, L.; You, Y.; Zou, X.; Chen, V.; Li, S.; Huang, G.; Hariharan, B.; Weinberger, K.Q. Resource aware person re-identification across multiple resolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8042–8051.
5. Huang, Y.; Zha, Z.J.; Fu, X.; Zhang, W. Illumination-invariant person re-identification. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 365–373.
6. Saquib Sarfraz, M.; Schumann, A.; Eberle, A.; Stiefelhagen, R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 420–429.
7. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep learning for person re-identification: A survey and outlook. *arXiv* **2020**, arXiv:2001.04193.
8. 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, 18–22 June 2018; pp. 79–88.
9. Xiao, T.; Li, S.; Wang, B.; Lin, L.; Wang, X. Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3415–3424.
10. Zhu, X.; Wu, B.; Huang, D.; Zheng, W.S. Fast open-world person re-identification. *IEEE Trans. Image Process.* **2017**, *27*, 2286–2300. [[CrossRef](#)] [[PubMed](#)]
11. Liu, Y.; Kong, L.; Chen, G.; Xu, F.; Wang, Z. Light-weight AI and IoT collaboration for surveillance video pre-processing. *J. Syst. Archit.* **2020**, 101934. [[CrossRef](#)]
12. Mittal, S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *J. Syst. Archit.* **2019**, *97*, 428–442. [[CrossRef](#)]
13. 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.

14. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 4510–4520.
15. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 6848–6856.
16. Wang, Y.; Chen, Z.; Wu, F.; Wang, G. Person re-identification with cascaded pairwise convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1470–1478.
17. 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, 26 June–1 July 2016; pp. 2818–2826.
18. 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, 18–22 June 2018; pp. 7132–7141.
19. Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. Abd-net: Attentive but diverse person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8351–8361.
20. Xu, S.; Cheng, Y.; Gu, K.; Yang, Y.; Chang, S.; Zhou, P. Jointly attentive spatial-temporal pooling networks for video-based person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4733–4742.
21. Chen, D.; Li, H.; Xiao, T.; Yi, S.; Wang, X. Video person re-identification with competitive snippet-similarity aggregation and co-attentive snippet embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1169–1178.
22. Fu, Y.; Wang, X.; Wei, Y.; Huang, T. Sta: Spatial-temporal attention for large-scale video-based person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February, 2019; Volume 33, pp. 8287–8294.
23. Zhang, Z.; Lan, C.; Zeng, W.; Chen, Z. Multi-Granularity Reference-Aided Attentive Feature Aggregation for Video-based Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 10407–10416.
24. 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.
25. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2285–2294.
26. Chang, X.; Hospedales, T.M.; Xiang, T. Multi-level factorisation net for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2109–2118.
27. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3702–3712.
28. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
29. Cubuk, E.D.; Zoph, B.; Mane, D.; Vasudevan, V.; Le, Q.V. Autoaugment: Learning augmentation policies from data. *arXiv* **2018**, arXiv:1805.09501.
30. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
31. 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; Springer: Cham, Switzerland, 2016; pp. 17–35.
32. Gray, D.; Tao, H. Viewpoint invariant pedestrian recognition with an ensemble of localized features. In Proceedings of the European Conference on Computer Vision, Marseille, France, 12–18 October 2008; pp. 262–275.
33. Loy, C.C.; Liu, C.; Gong, S. Person re-identification by manifold ranking. In Proceedings of the 2013 IEEE International Conference on Image Processing, Melbourne, Australia, 15–20 September 2013; pp. 3567–3571.
34. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 2014 IEEE 22nd International Conference on Pattern Recognition, Stockholm, Sweden, 24–28 August, 2014; pp. 34–39.
35. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
36. Lin, Y.; Zheng, L.; Zheng, Z.; Wu, Y.; Hu, Z.; Yan, C.; Yang, Y. Improving person re-identification by attribute and identity learning. *Pattern Recognit.* **2019**, *95*, 151–161. [[CrossRef](#)]
37. Bai, S.; Tang, P.; Torr, P.H.; Latecki, L.J. Re-ranking via metric fusion for object retrieval and person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 740–749.
38. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.

39. Chen, H.; Wang, Y.; Xu, C.; Shi, B.; Xu, C.; Tian, Q.; Xu, C. AdderNet: Do we really need multiplications in deep learning? In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 1468–1477.
40. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7036–7045.
41. Tan, M.; Le, Q.V. Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
42. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Online, 14–19 June 2020; pp. 10781–10790.
43. Redmon, J.; Farhadi, A. Yolo3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
44. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
45. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
46. Chen, L.C.; Yang, Y.; Wang, J.; Xu, W.; Yuille, A.L. Attention to scale: Scale-aware semantic image segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 3640–3649.
47. Yang, S.; Peng, G. Attention to Refine Through Multi Scales for Semantic Segmentation. In Proceedings of the Pacific Rim Conference on Multimedia, Hefei, China, September 21–22 2018; pp. 232–241.
48. Tao, A.; Sapra, K.; Catanzaro, B. Hierarchical Multi-Scale Attention for Semantic Segmentation. *arXiv* **2020**, arXiv:2005.10821.
49. Qian, X.; Fu, Y.; Jiang, Y.G.; Xiang, T.; Xue, X. Multi-scale deep learning architectures for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5399–5408.
50. Zhou, K.; Xiang, T. Torchreid: A library for deep learning person re-identification in pytorch. *arXiv* **2019**, arXiv:1910.10093.
51. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Mu Lee, K. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 402–419.
52. Shen, Y.; Li, H.; Xiao, T.; Yi, S.; Chen, D.; Wang, X. Deep group-shuffling random walk for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2265–2274.
53. Chen, D.; Xu, D.; Li, H.; Sebe, N.; Wang, X. Group consistent similarity learning via deep crf for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8649–8658.
54. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Interaction-and-aggregation network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 9317–9326.
55. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 2138–2147.
56. Li, W.; Zhu, X.; Gong, S. Person re-identification by deep joint learning of multi-loss classification. *arXiv* **2017**, arXiv:1705.04724.
57. Liu, X.; Zhao, H.; Tian, M.; Sheng, L.; Shao, J.; Yi, S.; Yan, J.; Wang, X. Hydraplus-net: Attentive deep features for pedestrian analysis. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 350–359.