electronics

MDPI

Article

An Efficient Multi-Branch Attention Network for Person Re-Identification

Ke Han [1], Mingming Zhu [1], Pengzhen Li [2], Jie Dong [1,*], Haoyang Xie [1] and Xiyan Zhang [1]

[1] School of Information Engineering, North China University of Water Resources and Electric Power, Zhengzhou 450046, China; hanke@ncwu.edu.cn (K.H.); z202210090905@stu.ncwu.edu.cn (M.Z.); xiehaoyang@ncwu.edu.cn (H.X.); zhangxiyan@stu.ncwu.edu.cn (X.Z.)
[2] Henan Institute of Geophysical Spatial Information Co., Ltd., Zhengzhou 450046, China; lipengzhen@hnmtxxzx.com
* Correspondence: dongjie@ncwu.edu.cn; Tel.: +86-13837199086

Abstract: Due to the absence of tailored designs that address challenges such as variations in scale, disparities in illumination, and instances of occlusion, the implementation of current person re-identification techniques remains challenging in practical applications. An Efficient Multi-Branch Attention Network over OSNet (EMANet) is proposed. The structure is composed of three parts, the global branch, relational branch, and global contrastive pooling branch, and corresponding features are obtained from different branches. With the attention mechanism, which focuses on important features, DAS attention evaluates the significance of learned features, awarding higher ratings to those that are deemed crucial and lower ratings to those that are considered distracting. This approach leads to an enhancement in identification accuracy by emphasizing important features while discounting the influence of distracting ones. Identity loss and adaptive sparse pairwise loss are used to efficiently facilitate the information interaction. In experiments on the Market-1501 mainstream dataset, EMANet exhibited high identification accuracies of 96.1% and 89.8% for Rank-1 and mAP, respectively. The results indicate the superiority and effectiveness of the proposed model.

Keywords: person re-identification; multi-feature fusion; attention mechanism; multi-branch network

Citation: Han, K.; Zhu, M.; Li, P.; Dong, J.; Xie, H.; Zhang, X. An Efficient Multi-Branch Attention Network for Person Re-Identification. *Electronics* **2024**, *13*, 3183. https://doi.org/10.3390/electronics13163183

Academic Editors: Haibin Wu, Aili Wang and Yuji Iwahori

Received: 6 July 2024
Revised: 4 August 2024
Accepted: 10 August 2024
Published: 12 August 2024

1. Introduction

With the continuous improvement of people's awareness of public safety, in order to maintain social security and prevent criminal behavior, intelligent monitoring technology has attracted wide attention from society. Person re-identification (Re-ID) is a fundamental task in computer vision. It aims to retrieve the same person from different cameras in a surveillance network. It generally involves feature extraction of the input image, distance metrics of the extracted features, and similarity ranking based on the distance value. When the distance between a query image and images in a gallery is relatively short, it signifies a high degree of similarity, significantly enhancing the likelihood that these images belong to the same individual. Therefore, extracting discriminative pedestrian image features plays a central role in person re-identification [1,2]. However, due to a series of problems such as different camera viewpoints, environmental noise interference, light changes, and changes in pedestrian posture, the enhancement of retrieval accuracy poses significant challenges.

Traditional person re-identification methods mainly focus on two aspects: Firstly, traditional person re-identification relies on hand-designed feature extraction. For example, the core idea of the HOG (histogram of oriented gradient) [3] method is to calculate and count the histogram of oriented gradient in the local area of the image to form the feature, which can quickly describe the local gradient feature of the object. The core idea of the SIFT (scale invariant feature transform) [4] method is to find key points in different scale spaces and calculate the direction of key points, so as to realize the scale invariant description of image features. The LOMO (local maximal occurrence) [5] method can effectively describe

the appearance information of pedestrians by calculating the color histogram of the image and combining the local maximal occurrence strategy. Secondly, traditional person re-identification relies on the similarity measure. For example, XQDA (cross-view quadratic discriminant analysis) [6] is used to learn the best similarity measure. Traditional methods have many limitations in extracting image information based on low-level visual features and cannot extract discriminative features in the face of complex and variable scenes of pedestrian images [7].

Although person re-identification has been studied in the academic community for many years, since 2016, with the successful application of deep learning in many fields, researchers have begun to try to apply deep learning to person re-identification [8]. With the advancement of deep learning, the early focus of Re-ID research has shifted towards developing robust feature representations to discern individual identities, while concurrently exploring effective distance metric techniques to establish image similarities. The feature representation of pedestrians can be divided into global features and local features. Global features refer to the feature extraction of the overall or global information of pedestrian images. This feature extraction process is not limited to a local region of the pedestrian but takes into account the whole range of the pedestrian. For the input image pair, Wang et al. [9] used two independent convolutional neural networks to extract the features of each image, respectively, combining the efficiency of single image feature extraction and the advantages of the Cross-image Information Extraction (CIR) method. In order to make the network learn more discriminative features, some studies add the attention mechanism. Song et al. [10] used the attention mechanism to separate the characters from the background in pedestrian images, and only extracted the pedestrian features in the image to avoid the noise introduced by the background. Although global feature learning has the advantage of being relatively simple, it does not always capture the specific regions of pedestrian images that are more discriminative. This means that the global features, although able to capture the overall information, may ignore those local details in the image that are crucial for distinguishing different pedestrians. Therefore, some s researchers have used local feature learning for person re-identification.

Local feature learning improves the model's robustness to locally misaligned scenes in pedestrian images by focusing on local regions of the image and learning the aggregated features of these regions. As a common local feature extraction method, image segmentation has been widely used in pedestrian recognition and other fields. Due to the particularity of person body structure, researchers usually divide images into several parts, such as head, upper body, legs, and feet, following the natural structure of the human body, in order to better capture and extract local features related to pedestrian identity. For example, Somers et al. [11] designed a body part attention module that uses external human semantic information to generate relevant local features. Sun et al. [12] proposed the PCB method to divide the pedestrian feature map into six equal blocks to learn local features and used the concatenated local features as feature descriptors. They also proposed the RPP method to adaptively partition image blocks according to the content similarity of each block. However, there are misaligned pedestrian images, so Zhang et al. designed a dynamic alignment network, AlignedReID [13], which has the ability to automatically and accurately align image blocks from top to bottom without relying on additional information. This automatic alignment mechanism enhances the robustness and performance of the person re-identification network. Pang et al. [14] designed the Generating Local Part (GLP) module to divide the feature map and generate features of occluded parts.

Recently, transformers have been widely used in nearly every computer vision scene. Dosovitskiy et al. [15] introduced the Transformer model into the field of image recognition for the first time and proposed the Vision Transformer (ViT) model. The Vision Transformer model formulates images as sequential data, subsequently inputting them into the Transformer model to accomplish various classification tasks. He et al. [16] proposed a Re-ID method, which integrates side information embeddings and a jigsaw patches module with a transformer to obtain robust features in a pure Transformer framework for improving

performance. Based on the ViT model, Heo et al. [17] proposed the Pooling-based Vision Transformer (PiT) model. The PiT combines pooling layers, which can reduce the large size of space in the ViT structure. The results show that after introducing the pooling layer, the spatial interaction area of Transformer becomes wider and the interaction rate is higher. Li et al. [18] utilized the disentangling capability of the Transformer model to propose the Part-aware Transformer, which is capable of decoupling robust features from distinct human body parts, making it suitable for the task of occluded person re-identification. Transformer-based methods often exhibit superior performance on extensive datasets, yet they typically necessitate greater computational resources and an abundant amount of training samples.

Aiming at the problems of environmental noise interference and the failure to take into account feature extraction at different scales and granularities in current person re-identification methods, inspired by the Relational network [2], this paper proposes an Efficient Multi-Branch Attention Network over OSNet [19]. Additionally, based on the OSNet backbone network, the attention mechanism module is introduced to refine the semantic expression of features, effectively improving the performance of the model. The correlation between features is increased by the combination of adjacent features. Finally, the multi-loss joint function is used to strengthen the supervised training of the model. The specific work of model construction and experimental deployment will be carried out in the following papers. Experimental results on widely used datasets demonstrate that the method proposed in this paper achieves impressive performance and robustness in Re-ID tasks. The primary contributions of this paper are as follows:

(1) This paper proposes an Efficient Multi-Branch Attention Network (EMANet), which comprises three branches: the global branch, relational branch, and global contrastive pooling branch.
(2) We introduce the DAS attention module, which focuses and increases attention to salient image regions, seamlessly integrating it into the backbone network to increase the accuracy of the person re-identification network.
(3) EMANet outperforms existing methods, achieving competitive results in popular benchmarks. Through rigorous ablation studies and visualization, we systematically validate the significance and contribution of each module and branch.

2. Methods

This section first introduces the overall network architecture and attention mechanism proposed in this paper, then introduces the global branch, relational branch, global contrastive pooling branch, and finally the loss functions.

2.1. Network Architecture

This paper proposes an Efficient Multi-Branch Attention Network over OSNet (EMANet). The overall network structure is shown in Figure 1. Initially, pedestrian images are preprocessed and resized to a uniform size of 256 $\times$ 128, which are then fed into the backbone network for training. Extensive experiments [20–22] have demonstrated that attention mechanisms significantly enhance the network's feature extraction capabilities. Consequently, DAS (Deformable Attention to Capture Salient Information) [23] is incorporated after the convolution of the third and fourth layers of OSNet. Compared to previous approaches for achieving attention in CNNs, DAS offers dense attention to the input features and examines the feature context in a holistic manner. Subsequently, following the fifth convolutional layer, the deep semantic features are routed into three separate branches: the global branch, relational branch, and global contrastive pooling branch. The multi-feature vectors F_1, F_2, and F_3 obtained from the three branches are fused into the feature vector F in the channel dimension. Finally, vector F is used as the basis for classification calculation.

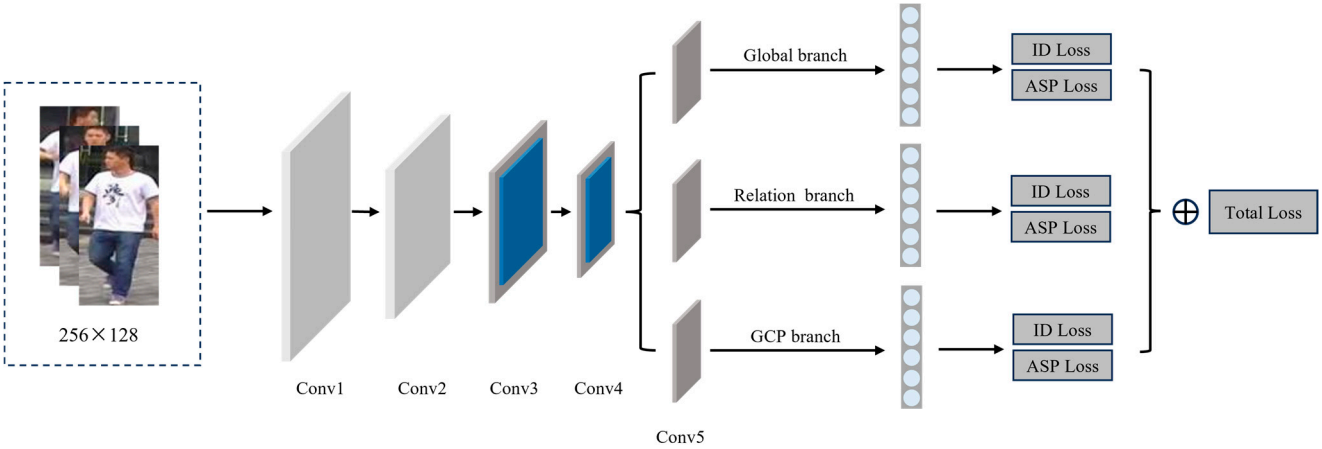


Figure 1. The overall architecture of EMANet. The network comprises three branches: the global branch, relational branch, and global contrastive pooling branch. The blue part represents the DAS.

2.2. Attention Mechanism Module

In order to improve the ability of the network to capture prominent features and make it focus more on the characteristics of pedestrians, an attention mechanism module is integrated into the backbone network. DAS attention combines Depthwise Separable Convolution (DSC) and Deformable Convolution (DC) to focus and increase attention on significant regions and compute the dense attention (pixel-wise) weights. It enhances the ability of convolutional networks to extract features in a computationally efficient way to provide focused attention on relevant content. The module is shown in Figure 2.

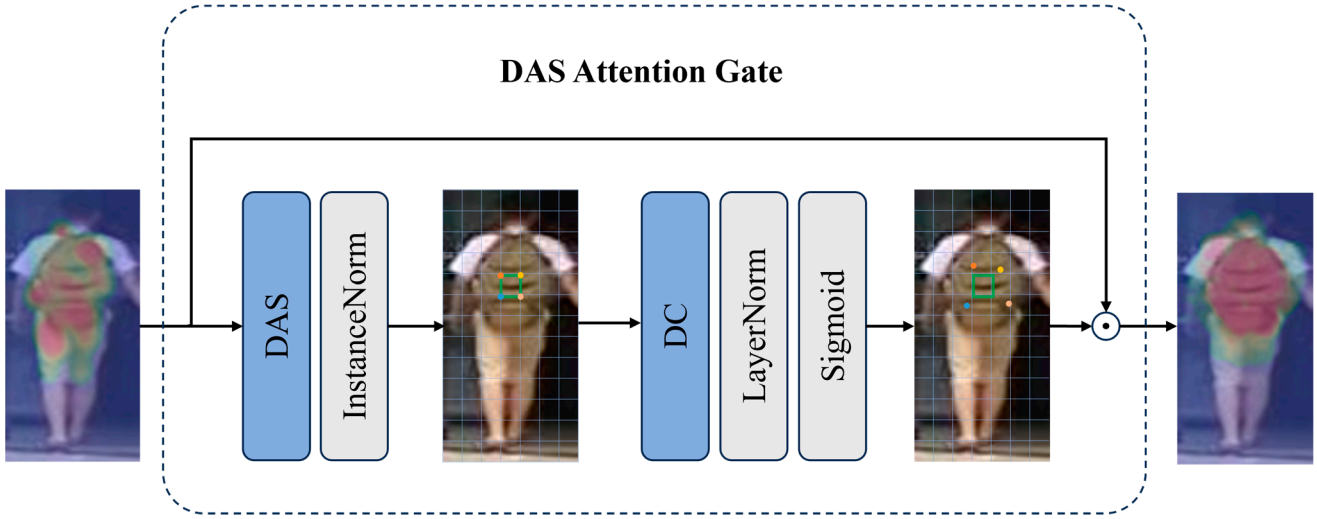


Figure 2. The leftmost heatmap depicts the saliency map generated by the OSNet without the integration of attention mechanism. In contrast, the rightmost heatmap showcases the same layer's activation pattern after the application of the DAS.

The number of channels of the feature is first reduced using the Depthwise Separable Convolution operation, which transforms the number of channels from c to $\alpha \times c$, where $0 < \alpha < 1$. The purpose of setting the size reduction hyperparameter α is to balance computational efficiency with accuracy. After reducing the number of channels, Instance Normalization is applied, followed by *GELU* nonlinear activation. These operations enhance the representativity of the features and contribute to the effectiveness of the attention mechanism. Equation (1) shows the compression process where X is the input

feature and W_1 represents the Depthwise Separable Convolution. Default α used in our implementation for this paper is 0.2.

$$X_c = GELU(InstanceNorm(XW_1)) \tag{1}$$

The input features are compressed by Equation (1) and then passed through a Deformable Convolution. Deformable Convolution adds 2D displacement to the mesh sampling positions of standard convolution rules, so that the sampling mesh can be freely deformed. Equation (2) shows the operation of Deformable Convolution, where k is the size of the kernel and its weights are w_k applied on the fixed reference points of p_{ref} the same way as regular kernels in CNNs. Δp is a trainable floating point parameter that helps the kernel to find the most relevant features. w_k is also trainable parameter between 0 and 1.

$$deform(p) = \sum_{k=1}^{K} w_k \cdot w_p \cdot X\Big(p_{ref,k} + \Delta p_k\Big) \tag{2}$$

Following the Deformable Convolution, we apply Layer Normalization, and then a Sigmoid activation function σ is used, as shown in Equation (3). The Deformable Convolution operation changes the number of feature channels from $\alpha \times c$ to the original input c.

$$A = \sigma(LayerNorm(deform(X_c))) \tag{3}$$

The attention tensor A obtained after Equation (3) represents the weight corresponding to each element of the original feature map. Each element in the tensor has a value between 0 and 1. These values determine which parts of the original feature map we emphasize or filter out. Finally, to incorporate the DAS mechanism into the CNN model, we perform pointwise multiplication between the original input tensor and the attention tensor obtained in the previous step. The symbol '$\odot$' denotes the pointwise multiplication operation.

$$X_{out} = X \odot A \tag{4}$$

The output of the multiplication in Equation (4) is directly utilized as input data for the subsequent layer of the CNN model, enabling a seamless integration with the attention mechanism without any adjustments to the backbone architecture.

2.3. Global Branch

The purpose of the global branch is to capture the overall features of the person image. Contrary to the majority of research that adopts ResNet [24] as the backbone network, our method selects OSNet as the backbone, with its architecture depicted in Figure 3. OSNet exhibits two notable advantages over ResNet. Firstly, it achieves light model weight by leveraging depthwise separable convolutions to significantly reduce the number of parameters. Secondly, it enhances the receptive field by extracting features from multiple scales, allowing for broader contextual understanding. OSNet introduces a Unified Aggregation Gate mechanism, which dynamically integrates multi-scale features based on the input. This mechanism enables the network to adaptively modulate the weights of features at different scales, thereby facilitating a more efficient utilization of multi-scale information and yielding discriminative global features. Consequently, the output of OSNet, without further specialized processing, serves as the global branch to obtain a $512 \times 1 \times 1$ feature vector F_1.

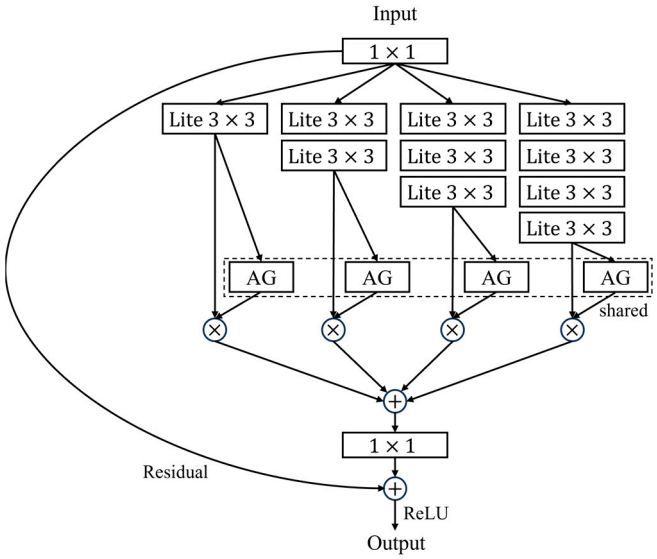


Figure 3. The architecture of OSNet. AG: Aggregation Gate.

2.4. One-vs.-Rest Relational Branch

Local features usually include the information from various regions of the feature map, yet they often represent only a fraction of the overall image and neglect the relationship between different body parts. To address this, the relational branch associates the local information with the corresponding remaining parts. As an illustration, Figure 4 depicts an example of extracting the local relational feature q_1.

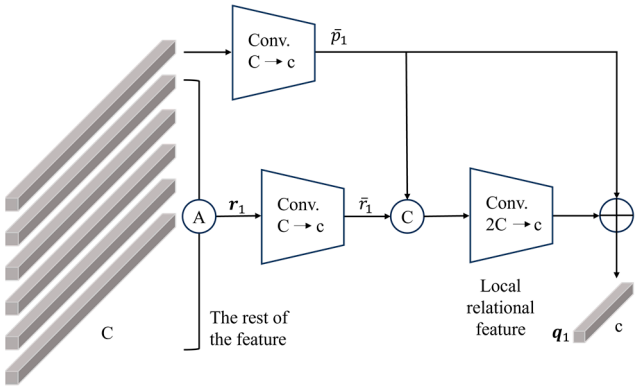


Figure 4. One-vs.-rest relational module.

Concretely, we denote by p_i $(i = 1, 2, \cdots, 6)$ each part-level feature of size $C \times 1 \times 1$. Take a local feature p_i as an example, and apply average pooling to the rest of the local specific to obtain r_i. The calculation formula is shown in Equation (5).

$$r_i = \frac{1}{5}\sum_{j \neq i} p_i, i, j = 1, 2, \cdots, 6 \tag{5}$$

We then add a 1×1 convolutional layer for each p_i and r_i, respectively, to obtain feature maps $\overline{p}_i$ and $\overline{r}_i$ of size $c \times 1 \times 1$. The two features $\overline{p}_i$ and $\overline{r}_i$ are concatenated and fed into a sub-network that performs dimensionality reduction from $2c$ to c, yielding the relational information between $\overline{p}_i$ and $\overline{r}_i$. The feature q_i contains information of the original one $\overline{p}_i$ itself and other body parts. Finally, skip connections are used to transfer the relational information of $\overline{p}_i$ and $\overline{r}_i$ to $\overline{p}_i$. The calculation formula is shown in Equation (6).

$$q_i = \overline{p}_i + R_p(T(\overline{p}_i, \overline{r}_i)), \ (i = 1, \ldots 6) \tag{6}$$

where T represents the concatenation of $\overline{p}_i$ and $\overline{r}_i$, forming a vector of size $2c$. R_p represents a sub-network consisting of a 1×1 convolution, batch normalization, and ReLU layers. The channel dimension is reduced from $2c$ to c via the sub-network operation.

2.5. Global Contrastive Pooling Branch

The global contrastive pooling branch can effectively mitigate the interference from background information, thereby enabling the extracted global features to be more concentrated on the pedestrian area. The GCP module is shown in Figure 5.

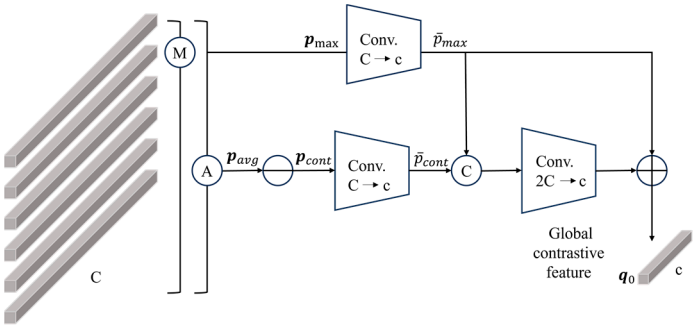


Figure 5. Global contrastive pooling module.

Initially, average and max pooling are performed on all part-level features. We denote the resulting feature maps obtained through average pooling and max pooling as p_{avg} and p_{max}, respectively. Subsequently, we compute a contrastive feature p_{cont} by subtracting p_{max} from p_{avg}, representing the discrepancy between them. It aggregates the most salient and discriminatory information from body parts except the one for p_{max}. We then add a 1×1 convolutional layer to reduce the number of channels of p_{cont} and p_{max} from C to c, denoted by $\overline{p}_{cont}$ and $\overline{p}_{max}$, respectively. The two features $\overline{p}_{cont}$ and $\overline{p}_{max}$ are concatenated and fed into a sub-network that performs dimensionality reduction from $2c$ to c, finally transferring the complementary information of the contrastive feature $\overline{p}_{cont}$ to $\overline{p}_{max}$. The number of channels of the output contrastive feature is consistent with $\overline{p}_{max}$. The calculation formula is shown in Equation (7).

$$q_0 = \overline{p}_{max} + R_g(T(\overline{p}_{max}, \overline{p}_{cont})) \tag{7}$$

where T represents the concatenation of $\overline{p}_i$ and $\overline{r}_i$, forming a vector of size $2c$. R_p represents a sub-network consisting of a 1×1 convolution, batch normalization, and ReLU layers. The channel dimension is reduced from $2c$ to c via the sub-network operation. Finally, the obtained feature q_0 is used as the output of the contrastive pooling branch, denoted as F_3.

2.6. Loss Functions

In this paper, we introduce two loss functions: the ID Loss and the adaptive sparse pairwise loss [25]. The network proposed in this paper adopts the above two loss functions to jointly supervise learning.

Label smooth cross entropy loss function is an improvement of cross entropy loss, which is used to optimize the pedestrian re-identification classification task. The loss function is calculated as follows:

Define the ID loss as follows:

$$L_{id} = -\frac{1}{N}\sum_{i=1}^{N} q_i \log(p_i) \tag{8}$$

where N represents the number of samples, p_i denotes the predicted probability for the ith identity, and q_i is the smooth label of identity i, which serves to prevent overfitting, which is defined as follows:

$$q_i = \begin{cases} 1 - \varepsilon\frac{N-1}{N}, & i = y \\ \frac{\varepsilon}{N}, & i \neq y \end{cases} \tag{9}$$

where y represents the person label and ε represents the error rate, which is used to reduce the confidence of the model on the labels of the training set and improve the generalization ability and is set to 0.1 in this paper.

Triplet Loss [26] and Circle Loss [27], which are widely used in person re-identification tasks, are based on a dense sampling mechanism, where each instance serves as an anchor to sample its positive and negative samples to form triplets. However, this mechanism inevitably introduces some positive pairs with minimal visual similarity, affecting the training effect. To address this, we propose using adaptive sparse pairwise loss, which selects only one hardest positive sample pair and one hardest negative sample pair for each class. The negative sample pair is the hardest negative sample pair between the class and all other classes, and the positive sample pair is the hardest positive sample pair in all sample sets. Adaptive sparse pairwise loss uses an adaptive positive mining strategy, which can dynamically adapt to different intra-class changes. The loss function is calculated as follows:

$$L_{sp} = \frac{1}{K}\sum_{i}^{K} \log\left(1 + e^{\frac{s_i^- - (\alpha_i S_{i,h}^+ + (1-\alpha_i) S_{i,lh}^+)}{\tau}}\right) \tag{10}$$

where K represents the total number of pedestrian categories in per mini-batch, τ is the temperature parameter, and S_i^- is denoted as the negative sample pair in the ith class. $S_{i,h}^+$ is denoted as a positive sample pair in the ith class. $S_{i,lh}^+$ is denoted as the least-hard positive. α_i is an adaptive weight to balance the hardest and the least-hard positive pairs for each class.

In order to improve the robustness of the network, ID Loss and adaptive sparse pairwise loss are combined and added to different stages of model training. The total loss function is the sum of the loss functions of each branch. The calculation formula is given in Equation (11), where λ is the balance parameter.

$$L_{sum} = L_{id} + \lambda L_{sp} \tag{11}$$

3. Experiments

Section 3.1 introduces the datasets and the evaluation protocols. Section 3.2 gives some implementation details. Section 3.3 compares our method to others. Section 3.4 provides ablation experiments to verify the effectiveness of the proposed method. Section 3.5 conducts numerous rigorous experiments to elucidate the impact of various parameters.

3.1. Datasets and Evaluation Protocols

In our experiments, we use three well-known image-based datasets: CUHK03 [28], DukeMTMC-reID [29], and Market-1501 [30].

CUHK03: CUHK03 is a dataset consisting of 1467 identities for 13,164 images. It is divided into two parts, CUHK-03 (labeled) with manually labeled pedestrian bounding boxes and CUHK-03 (detected) with DPM detected pedestrian bounding boxes.

DukeMTMC-reID: DukeMTMC-reID contains manually annotated boxes generated by eight cameras. It is composed by 36,411 images of 1404 identities. There are 16,522 images of 702 identities in the training set. The query and testing sets have 2228 and 17,661 images of 702 identities, respectively. For each identity in each camera, one image was selected as the query set in the test set, while the rest of the images were reserved as the image library.

Market-1501: Market-1501 has 32,668 images of 1501 person identities automatically detected from six disjoint cameras. The training set consists of 12,936 images of 751 identities.

The query set has 3368 probe images of 750 identities and the gallery set has 19,732 images with 750 identities.

In the task of person re-identification, two commonly utilized metrics, namely, Cumulative Matching Characteristic (CMC) and mean Average Precision (mAP), are employed to assess the performance. The traditional accuracy indicator, commonly known as Rank-1 accuracy, serves as a metric to quantify the congruency between the identity predicted by the model with the highest probability and the corresponding ground truth. Conversely, Rank-5 accuracy offers an alternative perspective, assessing the accuracy by considering the identities predicted by the model with the five highest probabilities.

3.2. Implementation Details

Our model was implemented using the PyTorch 1.12.1. We conducted experiments with an NVIDIA GTX 3090 GPU, utilizing OSNet as the backbone network and employing the weights pre-trained on ImageNet for fine-tuning. All images were resized to 256×128. The data augmentation included random flip and random erasing [31] with a probability of 50%. There were 64 images per mini-batch and four images per person ID. We employed the AMSGrad [32], with a momentum of 0.9 and an initial learning rate of 0.0015. The network was trained for 150 epochs. Additionally, we implemented a learning rate decay strategy, where the learning rate was reduced by a factor of 0.1 every 60 epochs.

3.3. Contrast Test

In order to verify the superiority of the person re-identification algorithm proposed in this paper, the EMANet is compared with some existing advanced person re-identification methods. The selected datasets for the experiment include CUHK03, DukeMTMC-reID, and Market-1501, and the experimental results do not adopt Re-ranking method. The performance comparison is shown in Table 1.

Table 1. Performance (%) comparison of different algorithm on CUHK03, DukeMTMC-reID, Market-1501 datasets. The best results of all experiments are shown in bold. The symbol '-' denotes that the relevant paper does not provide data.

Metheds	CUHK03-Labeled		CUHK03-Detected		DukeMTMC-reID		Market-1501	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
OSNet	-	-	72.3	67.8	88.6	73.5	94.8	84.8
SVDNet [33]	40.9	37.8	41.5	37.3	76.7	56.8	82.3	62.1
Pyramid [34]	78.9	76.9	78.9	74.4	89.0	79.0	**96.7**	88.2
AlignedReID	-	-	61.5	59.6	69.7	82.1	91.8	79.1
DGNet [35]	-	-	-	-	86.6	74.8	94.8	86.0
BDB [36]	79.4	76.7	76.4	73.5	89.0	76.0	95.3	86.7
CASN [37]	73.7	68.0	71.5	64.4	87.7	73.7	94.4	82.8
FED [21]	-	-	-	-	89.4	78.0	95.0	86.3
SCS+ [38]	80.3	77.2	77.1	74.3	90.3	80.9	96.0	89.4
With Res2Net50 [39]	-	-	-	-	88.1	77.6	95.0	87.1
UV-ReID-ABLM [40]	-	-	-	-	81.4	61.8	89.9	75.0
ICAM [41]	-	-	-	-	85.6	71.6	93.3	82.3
PSF-C-Net [42]	-	-	-	-	87.1	76.9	95.2	87.3
EMANet	**88.1**	**89.4**	**83.6**	**85.4**	**90.6**	**81.0**	96.1	**89.8**

From Table 1, it is observed that EMANet outperforms most mainstream methods, achieving the result in both Rank-1 of 96.1% and the mAP of 89.8% on Market-1501, which exhibits excellent performance. EMANet achieves the best results in both Rank-1 of 90.6%

and the mAP of 81.0% on DukeMTMC-reID, surpassing the second-best method by 0.3% and 0.1%, respectively. Similarly, on CUHK03-Labeled, our method achieves the best Rank-1 of 88.1% and the mAP of 89.4%, surpassing the second best method by 8.7% and 12.2%. Specifically, compared to the benchmark network OSNet, EMANet demonstrates improvements of 11.3% and 17.6% in Rank-1 and mAP on the CUHK03-Detected. In comparison to SVDNet, which based on global feature learning, EMANet improves Rank-1 and mAP by 13.8% and 27.7% on Market-1501, respectively. In comparison to Pyramid for multi-scale feature extraction, EMANet improves Rank-1 and mAP on DukeMTMC-reID by 1.6% and 2.0%, respectively. This suggests that EMANet can obtain discriminative feature representation through multi-scale feature extraction.

3.4. Ablation Experiment

In order to verify the effectiveness of the different branches and DAS attention mechanism, ablation experiments were performed using the DukeMTMC-reID dataset. In the experiment, Baseline is the network proposed in the literature [19], where Attention represents the DAS, GCP represents the global contrastive pooling branch, and Re represents the relational branch. The results of ablation experiments are shown in Table 2. After adding the GCP and Re branch, the mAP and Rank-1 are improved by 0.7% and 3.9%, respectively. Utilizing the combined application of the attention mechanism, GCP, and Re branch, we observe a significant enhancement. Specifically, the Rank-1 accuracy has increased by 2.0%, while mAP has improved by 7.5%. The experimental results show that adding branches to a certain extent can improve the performance of the network, and multiple branches play a complementary role in the network structure.

Table 2. Results of each module added on the DukeMTMC-reID dataset. The best results are shown in bold.

Method	Rank-1	mAP
BaseLine	88.6	73.5
BaseLine + GCP	89.0	74.2
BaseLine + Re	89.6	75.5
BaseLine + GCP + Re	89.3	77.4
BaseLine + Attention + GCP + Re	**90.6**	**81.0**

In order to further verify that our method has a strong ability to extract pedestrian features, a heatmap based on different branches is displayed in Figure 6. Specifically, Figure 6a represents the original pedestrian image, Figure 6b depicts the heatmap activation of the baseline network, Figure 6c shows the heatmap after the addition of the attention mechanism, and Figure 6d displays the heatmap corresponding to the inclusion of the relational module. Figure 6e presents the heatmap resulting from the addition of the GCP module. The red areas represent the focal regions attended to by the models, whereas the blue areas denote less significant regions. As seen in Figure 6, after integrating the attention mechanism into the benchmark network, the pedestrian area extracted by the network becomes larger and is concentrated on distinguishable parts. The heatmap activation demonstrates that the relational module enables each horizontal local area to incorporate information from the remaining horizontal local areas. Subsequently, with the addition of the GCP module, the feature map of the whole body can be extracted. The ablation experiments demonstrate the effectiveness of the multi-branch module, and the interplay between the various branches enhances the accuracy of the person re-identification model.

In our experimental analysis presented in Table 3, we conduct experiments to analyze the influences of DAS in different layers. As shown in Table 3, when selecting a single layer for DAS integration, Layers 3 and 4 exhibit promising performance gains. When inserting DAS in multiple layers, Layer 34 obtains the best performance. In the context of our experimental results, the performance of Layer 234 exhibits a marginal decrease compared to that of Layer 34. This slight decrement in performance may be attributed to the

fact that the feature representations at the second layer are not yet sufficiently mature and stable, thereby limiting the optimal functioning of the attention mechanism. Conversely, by incorporating the DAS at Layer 34, the model is able to leverage the more stable and meaningful features extracted from preceding layers, enabling the attention mechanism to focus on key information more effectively.

(**a**) (**b**) (**c**) (**d**) (**e**)

Figure 6. (**a**) The original pedestrian image; (**b**) The heatmap activation of the baseline network; (**c**) The heatmap activation after the addition of the attention mechanism; (**d**) The heatmap activation of the relational module; (**e**) The heatmap activation of the GCP module.

Table 3. The performances of different layers to place DAS on the Market1501 dataset. The best results are shown in bold.

Method	mAP	Rank-1	Rank-5	Rank-10
Layer 1	87.8	94.3	97.9	89.8
Layer 2	89.1	94.8	98.1	89.7
Layer 3	88.6	95.2	**98.5**	89.6
Layer 4	89.7	95.1	98.1	**99.3**
Layer 12	88.9	95.6	97.8	98.1
Layer 23	89.5	95.9	98.2	98.7
Layer 34	**89.8**	**96.1**	**98.5**	**99.3**
Layer 234	89.3	95.6	98.2	99.0

3.5. Parameters Analysis

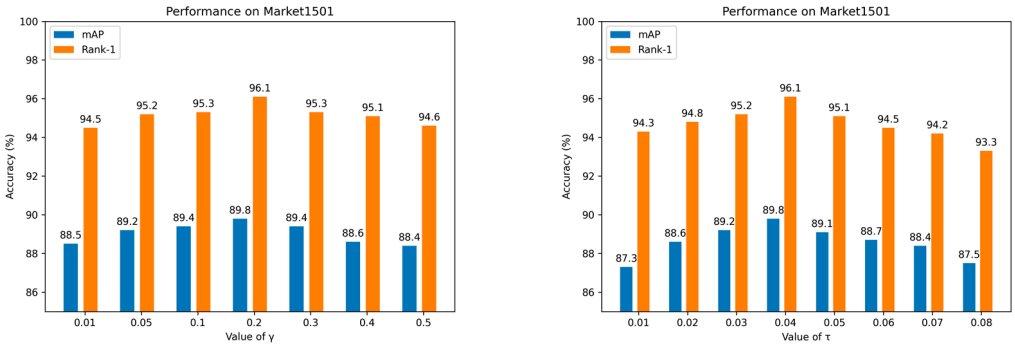
In this section, we conduct a series of rigorous experiments on the Market1501 dataset to elucidate the impact of various parameters on the performance of our method.

Parameter analysis of λ: The parameter λ represents the weight of the ASP loss in Equation (11). In our experiments, the λ parameter was varied from 0.001 to 0.5. The results of these experiments are presented in Figure 7a. It can be observed that the overall performance of the network is not overly sensitive to the variation of the λ parameter. Specifically, when the value of λ is set to 0.2, the network achieves the best performance.

Parameter analysis of τ: The Parameter τ is the temperature parameter in the ASP loss. The parameter τ was varied from 0.01 to 0.08 with an interval of 0.01. The experimental results are shown in the Figure 7b. It is particularly important to note that when temperature τ is 0.04, we method achieves the best Rank-1.

Numbers of grids: In the One-vs.-rest relational branch and global contrastive pooling branch, the feature maps are segmented horizontally into different amounts, such as q^{P_2} and q^{P_4}, and represent splitting the feature map into two and four horizontal regions, respectively. It is noteworthy that the variables q^{P_2}, q^{P_4}, and q^{P_6} embody distinct local relational characteristics, thereby exhibiting diverse global contrastive features. We conducted experiments on the DukeMTMC-reID and Market1501 datasets to find the best number

of grids, and the results are shown in Table 4. We can see that the model shows the best performance when the number of grids is 6.



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

Figure 7. (**a**) Influence of λ on mAP and Rank-1 on Market1501; (**b**) influence of τ on mAP and Rank-1 on Market1501.

Table 4. Comparison between q^{P_2}, q^{P_4}, and q^{P_6}.

Amount	DukeMTMC-reID		Market-1501	
	Ranl-1	**mAP**	**Ranl-1**	**mAP**
q^{P_2}	88.7	79.3	94.1	88.3
q^{P_4}	88.6	80.4	95.2	88.4
q^{P_6}	90.6	81.0	96.1	89.8

After conducting a thorough parameter study, it was observed that the highest mAP is attained at a parameter λ setting of 0.2, a parameter τ of 0.2, and a horizontal division of 6 blocks. Based on these findings, we adopted the same optimal hyperparameters for the entirety of the experimental procedures conducted in this study.

3.6. Inference Time Analysis

In order to evaluate the impact of three branches on feature extraction time, we conducted an attention-based experiment. The inference time after adding different branches is listed in Table 5. Therein, feature extraction time refers to the time required for each 64 images processed in the inference process.

Table 5. Inference times of three branches for the DukeMTMC-reID and Market-1501 datasets.

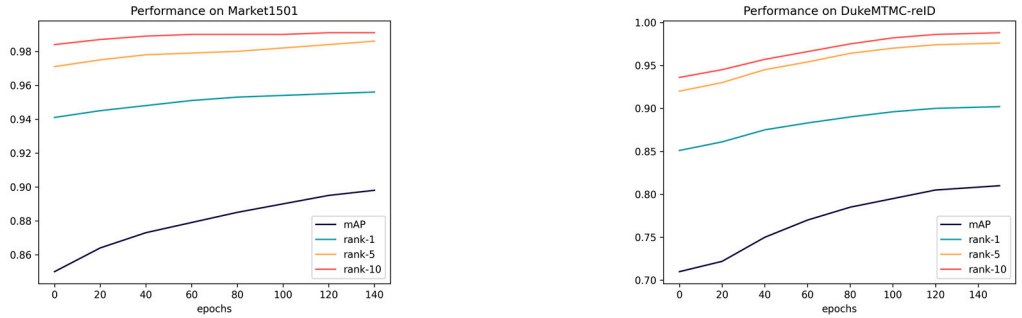
Title 1	DukeMTMC-reID	Market-1501
BaseLine + Attention	0.2134 s	0.1947 s
BaseLine + Attention + GCP	0.2253 s	0.2183 s
BaseLine + Attention + Re	0.2545 s	0.2337 s
BaseLine + Attention + GCP + Re	0.2845 s	0.2786 s

It can be observed from the table that the inference time of the network increases with increasing branches. The increase in time means that the network takes extra time to extract features of different expressivity.

4. Results Visualization

To demonstrate the effectiveness of the proposed algorithm in a more intuitive manner, Figure 8 shows the trends of Rank-1, Rank-5, Rank-10, and mAP during a single training process of the proposed network. As can be observed from Figure 8, all metrics gradually increase with the increasing number of epochs. Specifically, during the first 40 epochs,

the upward trend is relatively significant, and by the 120th epoch, the various evaluation metrics have gradually exhibited a stable pattern. Therefore, in this study, a total of 150 epochs was chosen for in-depth analysis.



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

Figure 8. (**a**) mAP and Rank-n change with epochs on Market1501; (**b**) mAP and Rank-n change with epochs on DukeMTMC-reID.

We compare the ranking results on the Market-1501 dataset with the results from other methods in Figures 9–11. Query refers to the query image, and the images marked 1 to 10 are the top ten results that are most relevant to the query image retrieved from the gallery library. The positive sample images, which are unlabeled, represent entities of the same pedestrian as the query. The negative sample images, annotated with red bounding boxes, represent entities of different pedestrians from the query. As can be seen from Figure 8, there are still negative samples in the retrieved pedestrian images, but most of them are correct pedestrian images in the retrieved results.

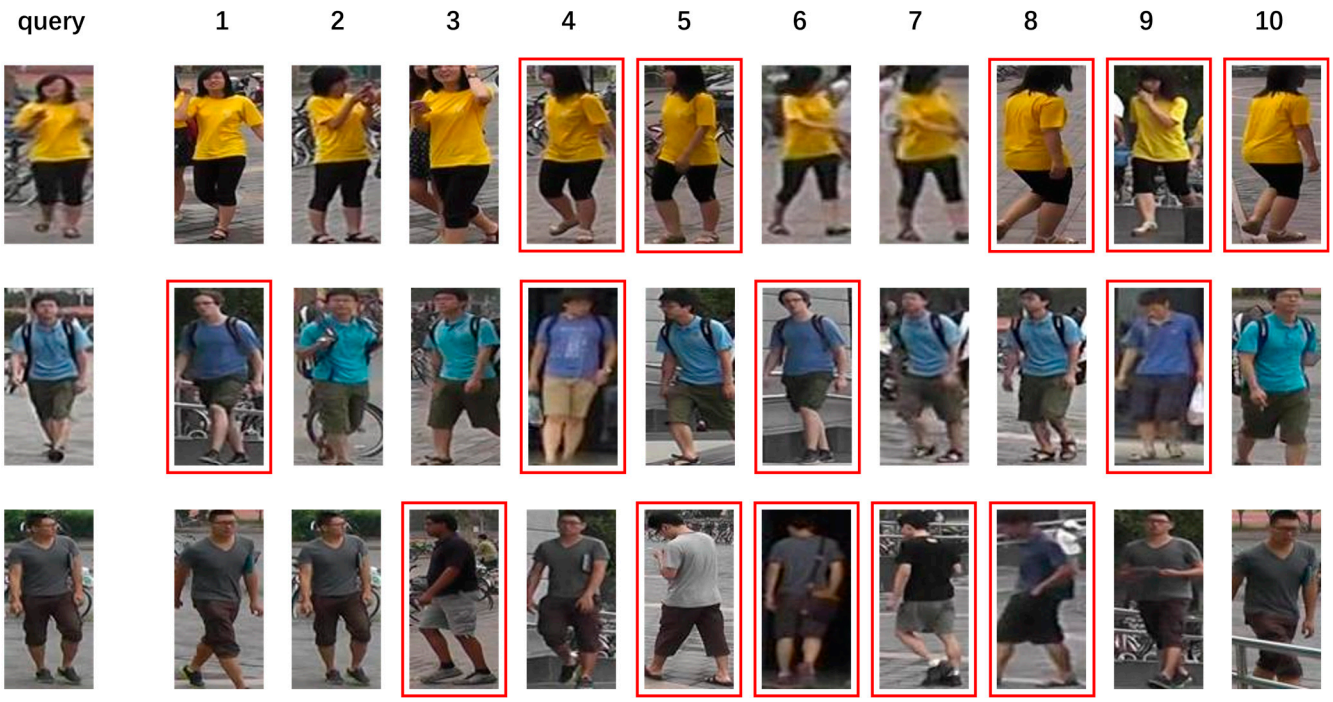


Figure 9. PCB visualized experimental results on Market-1501.

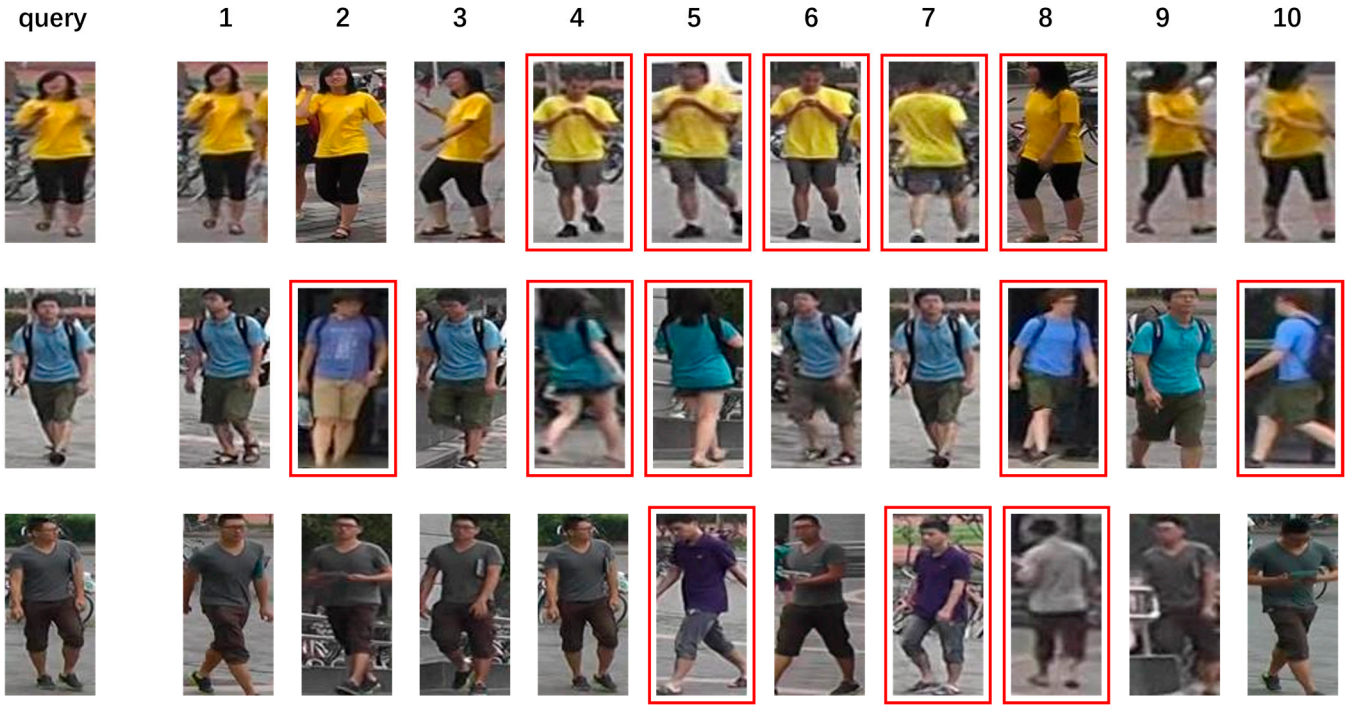


Figure 10. AlignedReID visualized experimental results on Market-1501.

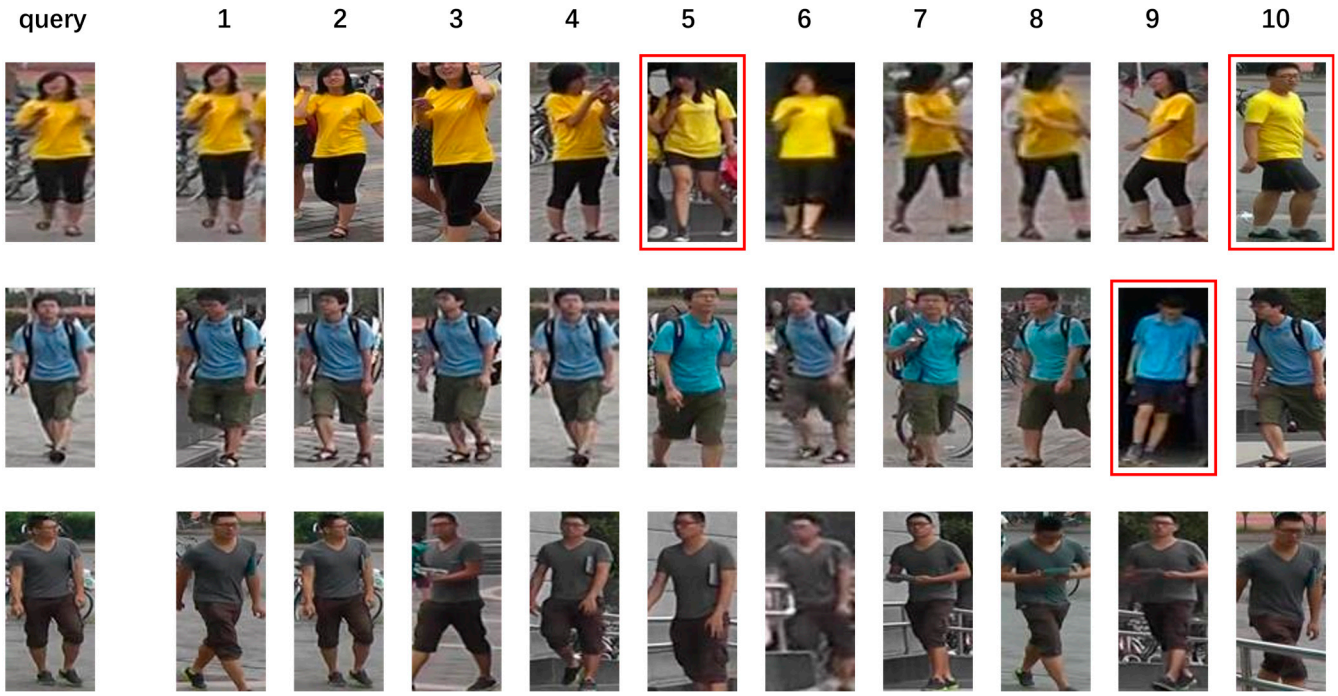


Figure 11. EMANet visualized experimental results on Market-1501.

5. Conclusions

In this paper, we propose an Efficient Multi-Branch Attention Network over OSNet (EMANet) for extracting discriminative pedestrian features. The EMANet primarily consists of three branches: a global branch, a relational branch, and a global contrastive pooling branch. The global branch extracts the overall information of pedestrians, while the relational branch captures the relationship between local features. The global contrastive pooling branch effectively removes interference from the background, ensuring that the extracted global features are more focused on the pedestrian area. By incorporating atten-

tion mechanisms into the OSNet backbone network, the model can automatically select and weight the importance of different features, enabling it to focus more on salient regions and features. The network is trained jointly using identity loss and adaptive sparse pairwise loss. Finally, the model is verified on three datasets. The experimental results demonstrate that the proposed method is effective for person re-identification tasks and worthy of reference.

Author Contributions: Conceptualization, K.H. and M.Z.; Investigation, P.L.; Methodology, K.H. and M.Z.; Software, M.Z.; Supervision, H.X.; Validation, H.X. and X.Z.; Visualization, X.Z. and J.D.; Writing—original draft, M.Z.; Writing—review and editing, M.Z. and H.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partly supported by the Research and Practice of Talent Cultivation Mode for Information Technology Innovation in Modern Industrial Colleges under the Background of New Engineering Education under Grant No. 2024SJGLX0108. This work was supported in part by the National Natural Science Foundation of China under Grant No. 82202270.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors upon request.

Conflicts of Interest: Author Pengzhen Li was also employed by Henan Institute of Geophysical Spatial Information Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial of financial relationships that could be construed as a potential conflict of interest.

References

1. Wu, C.; Ge, W.; Wu, A.; Chang, X. Camera-conditioned stable feature generation for isolated camera supervised person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 20238–20248.
2. Park, H.; Ham, B. Relation network for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 11839–11847.
3. Navneet, D. Histograms of oriented gradients for human detection. In Proceedings of the International Conference on Computer Vision & Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; Volume 2, pp. 886–893.
4. Lowe, D.G. Object recognition from local scale-invariant features. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–27 September 1999; Volume 2, pp. 1150–1157.
5. Liao, S.; Hu, Y.; Zhu, X.; Li, S.Z. Person re-identification by local maximal occurrence representation and metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2197–2206.
6. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
7. Wang, J.; Wang, J. MHDNet: A Multi-Scale Hybrid Deep Learning Model for Person Re-Identification. *Electronics* **2024**, *13*, 1435. [CrossRef]
8. Xu, D.; Chen, J.; Chai, X. An Orientation-Aware Attention Network for Person Re-Identification. *Electronics* **2024**, *13*, 910. [CrossRef]
9. Wang, F.; Zuo, W.; Lin, L.; Zhang, D.; Zhang, L. Joint learning of single-image and cross-image representations for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1288–1296.
10. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model 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. 1179–1188.
11. Somers, V.; De Vleeschouwer, C.; Alahi, A. Body part-based representation learning for occluded person re-identification. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 2–7 January 2023; pp. 1613–1623.
12. 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.
13. Zhang, X.; Luo, H.; Fan, X.; Xiang, W.; Sun, Y.; Xiao, Q.; Jiang, W.; Zhang, C.; Sun, J. Alignedreid: Surpassing human-level performance in person re-identification. *arXiv* **2017**, arXiv:1711.08184.

14. Pang, Y.; Zhang, H.; Zhu, L.; Liu, D.; Liu, L. Feature generation based on relation learning and image partition for occluded 147person re-identification. *J. Vis. Commun. Image Represent.* **2023**, *91*, 103772. [CrossRef]
15. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
16. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 15013–15022.
17. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021; pp. 11936–11945.
18. Li, Y.; He, J.; Zhang, T.; Liu, X.; Zhang, Y.; Wu, F. Diverse part discovery: Occluded person re-identification with part-aware transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 11–17 October 2021; pp. 2898–2907.
19. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale feature learning for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.
20. Zhou, J.; Dong, Q.; Zhang, Z.; Liu, S.; Durrani, T.S. Cross-modality person re-identification via local paired graph attention network. *Sensors* **2023**, *23*, 4011. [CrossRef] [PubMed]
21. Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; Song, J. Feature erasing and diffusion network for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4754–4763.
22. Chen, Y.; Wang, H.; Sun, X.; Fan, B.; Tang, C.; Zeng, H. Deep attention aware feature learning for person re-identification. *Pattern Recognit.* **2022**, *126*, 108567. [CrossRef]
23. Salajegheh, F.; Asadi, N.; Saryazdi, S.; Mudur, S. DAS: A Deformable Attention to Capture Salient Information in CNNs. *arXiv* **2023**, arXiv:2311.12091.
24. 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, 27–30 June 2016; pp. 770–778.
25. Zhou, X.; Zhong, Y.; Cheng, Z.; Liang, F.; Ma, L. Adaptive sparse pairwise loss for object re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19691–19701.
26. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
27. Sun, Y.; Cheng, C.; Zhang, Y.; Zhang, C.; Zheng, L.; Wang, Z.; Wei, Y. Circle loss: A unified perspective of pair similarity optimization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6398–6407.
28. Li, W.; Zhao, R.; Xiao, T.; Wang, X. Deepreid: Deep filter pairing neural network for person re-identification. In Proceedings of the IEEE Conference on computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 152–159.
29. 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, 11–14 October 2016; pp. 17–35.
30. 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.
31. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random erasing data augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008.
32. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237.
33. Sun, Y.; Zheng, L.; Deng, W.; Wang, S. Svdnet for pedestrian retrieval. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3800–3808.
34. Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; Ji, R. Pyramidal person re-identification via multi-loss dynamic training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8514–8522.
35. 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/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9317–9326.
36. Dai, Z.; Chen, M.; Gu, X.; Zhu, S.; Tan, P. Batch dropblock network for person re-identification and beyond. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3691–3701.
37. Zheng, M.; Karanam, S.; Wu, Z.; Radke, R.J. Re-identification with consistent attentive siamese networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5735–5744.
38. Zhu, K.; Guo, H.; Liu, S.; Wang, J.; Tang, M. Learning semantics-consistent stripes with self-refinement for person re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8531–8542. [CrossRef] [PubMed]
39. Mamedov, T.; Kuplyakov, D.; Konushin, A. Approaches to Improve the Quality of Person Re-Identification for Practical Use. *Sensors* **2023**, *23*, 7382. [CrossRef] [PubMed]
40. Perwaiz, N.; Shahzad, M.; Fraz, M. Ubiquitous vision of transformers for person re-identification. *Mach. Vis. Appl.* **2023**, *34*, 27. [CrossRef]

41. Wang, M.; Ma, H.; Huang, Y. Information complementary attention-based multidimension feature learning for person re-identification. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106348. [CrossRef]
42. Sun, R.; Chen, Q.; Dong, H.; Zhang, H.; Wang, M. PSF-C-Net: A Counterfactual Deep Learning Model for Person Re-Identification Based on Random Cropping Patch and Shuffling Filling. *Mathematics* **2024**, *12*, 1957. [CrossRef]