

Article

A Person Re-Identification Method Based on Multi-Branch Feature Fusion

Xuefang Wang¹, Xintong Hu^{2,*}, Peishun Liu^{2,*}  and Ruichun Tang²
¹ School of Mathematical Sciences, Ocean University of China, Qingdao 266100, China; wangxuefang@ouc.edu.cn

² Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266100, China; tangruichun@ouc.edu.cn

* Correspondence: huxintong@stu.ouc.edu.cn (X.H.); liups@ouc.edu.cn (P.L.)

Abstract: Due to the lack of a specific design for scenarios such as scale change, illumination difference, and occlusion, current person re-identification methods are difficult to put into practice. A Multi-Branch Feature Fusion Network (MFFNet) is proposed, and Shallow Feature Extraction (SFF) and Multi-scale Feature Fusion (MFF) are utilized to obtain robust global feature representations while leveraging the Hybrid Attention Module (HAM) and Anti-erasure Federated Block Network (AFBN) to solve the problems of scale change, illumination difference and occlusion in scenes. Finally, multiple loss functions are used to efficiently converge the model parameters and enhance the information interaction between the branches. The experimental results show that our method achieves significant improvements over Market-1501, DukeMTMC-reID, and MSMT17. Especially on the MSMT17 dataset, which is close to real-world scenarios, MFFNet improves by 1.3 and 1.8% on Rank-1 and mAP, respectively.

Keywords: intelligent video analysis; deep learning; model recognition; person re-identification method; Multi-Branch Feature Fusion Network



Citation: Wang, X.; Hu, X.; Liu, P.; Tang, R. A Person Re-Identification Method Based on Multi-Branch Feature Fusion. *Appl. Sci.* **2023**, *13*, 11707. <https://doi.org/10.3390/app132111707>

Academic Editor: Douglas O'Shaughnessy

Received: 6 September 2023

Revised: 21 October 2023

Accepted: 23 October 2023

Published: 26 October 2023



Copyright: © 2023 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 (Person re-ID) is of significant importance in video understanding. It aims to automatically identify pedestrian images in video images of different scenes based on a given picture of a target pedestrian. With the rapid development of deep learning, person re-ID methods based on deep learning are constantly emerging.

These methods can extract advanced abstract features and effectively improve the accuracy of model recognition. The methods for extracting features from pedestrian images can be categorized into two distinct groups: global feature learning methods and local feature learning methods. Global feature learning methods usually extract features from pedestrian images [1–3], which makes it difficult to capture detailed information about pedestrians. Local feature-learning methods aim to learn the discriminative features of pedestrians and ensure the alignment of each local feature. Commonly, local feature learning methods include predefined stripe segmentation [4,5], multi-scale fusion [6–8], soft attention [9–11], pedestrian semantic extraction [11–13], and global-local feature-learning [14–16]. These techniques can mitigate challenges, such as occlusions, errors in boundary detection, and variations in viewpoints and poses.

Local feature learning methods excel at capturing intricate object details, yet their reliability can be compromised by variations in posture and occlusion. Consequently, prior studies have amalgamated fine-grained local features with grained global features to augment the representational capacity. Wang et al. [17] proposed a multi-granularity feature learning strategy with global and local information, including one global feature learning branch to learn coarse-grained features and two local feature learning branches to learn fine-grained features. Ming et al. [18] designed a Global-Local Dynamic Feature Alignment

Network (GLDFA-Net) framework with global and local branches and proposed a local sliding alignment method into the local branch to calculate the distance measure. However, the feature extraction networks of these models still lack specific designs for scenarios, such as scale changes [7,19,20], illumination differences [21,22], and occlusions [23–25], so it is hard to apply them to real scenarios.

In previous studies, particular attention was given to these common issues. Wang et al. [25] proposed a novel Feature Erasing and Diffusion Network (FED), which incorporates an occlusion removal module trained to remove NPOs (Non-Pedestrian Occlusions) features based on predicted occlusion scores, guided by an image-level NPOs enhancement strategy. Subsequently, the feature diffusion module facilitates feature diffusion between the NPOs feature-erased pedestrian representation and memory features to synthesize Non-Target Pedestrian (NTP) features in the feature space. Liao et al. [7] introduced a person re-ID method that, under a saliency model, enabling processing of images and extraction of salient area images of human bodies in pedestrian images to mitigate interference from complex backgrounds for person re-ID tasks. However, there is limited research addressing joint solutions to multiple typical problems.

To address the aforementioned challenges, this paper proposes a Multi-Branch Feature Fusion Network (MFFNet) for addressing person re-ID problems in diverse scenarios. MFFNet comprises three branches: the multi-scale fusion global branch, the hybrid attention local branch, and the saliency guidance local branch. The global branch employs multi-scale feature fusion modules to balance the relationship between shallow image texture features and deep image semantic features, thereby enhancing the ability of the model to extract pedestrian features at various scales, scale changes, and illumination differences. The hybrid attention local branch utilizes HAM to focus on key information within spatial and channel features. Additionally, AFBN is employed to tackle occlusions by guiding the model to concentrate on the salient and sub-salient regions of pedestrian image features. Experimental results demonstrate the effectiveness of our approach with Rank-1 of 95.2%, 92.5%, and 80.2% achieved on the Market1501, DukeMTMC-reID, and MSMT17 datasets, respectively, outperforming other methods in [22,25–34].

2. Related Work

2.1. Transformer in Vision

Dosovitskiy et al. [35] introduced the Transformer model into the field of image recognition for the first time and proposed the Vision Transformer (ViT) model, which represents images as sequence data and inputs them into the Transformer model to achieve classification tasks. Subsequently, Touvron et al. [36] introduced knowledge distillation and data augmentation strategies, as well as efficient training strategies, which made ViT achieve a significant improvement in data efficiency. Liu et al. [37] proposed the Swin Transformer model, which effectively captures image features of different scales by introducing a hierarchical structure based on shifted windows and obtains global context information through hierarchical interaction. This method reduces the computational complexity and further improves the representation and visual perception ability of ViT in processing large-size images. To facilitate the extraction of features containing multi-level semantic information, we employ the Swin Transformer model in conjunction with shallow feature extraction and multi-scale feature fusion.

2.2. Attention Mechanism

The attention mechanism can aid the model in grasping the key information of the data to improve the ability of the model to extract essential features. Jie et al. [38] proposed a “Squeeze-and-Excitation (SE)” architecture, which can help the model achieve dynamic correction of image channel features. Li et al. [39] proposed a Selective Kernel (SK) unit that can gather information from multiple kernels, which enables neurons to dynamically adjust the receptive field size based on the input information. The above methods apply channel attention and spatial attention separately, ignoring the influence and correlation between

them. Therefore, Woo et al. [40] proposed a Convolutional Block Attention Module (CBAM), which blends channel attention and spatial attention, resulting in notable performance enhancements while minimizing computational overhead.

3. Multi-Branch Feature Fusion Network

To address the challenges of person re-ID in complex scenes, we employ a “scene decomposition” strategy and introduce the Multi-Branch Feature Fusion Network (MFFNet) in this section. Specifically, the MFFNet contains three different branch networks: a multi-scale fusion global branch (Section 3.1), a hybrid attention local branch (Section 3.2), and a saliency guidance local branch (Section 3.3). The overall structure of the model is shown in Figure 1.

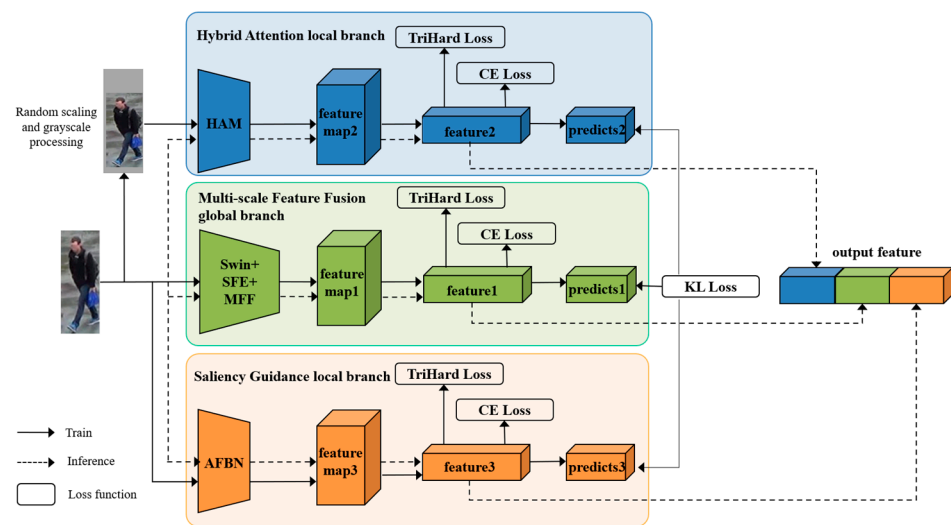


Figure 1. The overall framework of MFFNet.

3.1. Multi-Scale Feature Fusion Global Branch

This branch is designed hierarchically in an end-to-end manner, utilizing a Swin Transformer [37] as the backbone network. It incorporates Shallow Feature Extraction (SFE) and Multi-scale Feature Fusion (MFF) techniques to extract diverse global feature information from pedestrians across multiple scales. The structure is depicted in Figure 2.

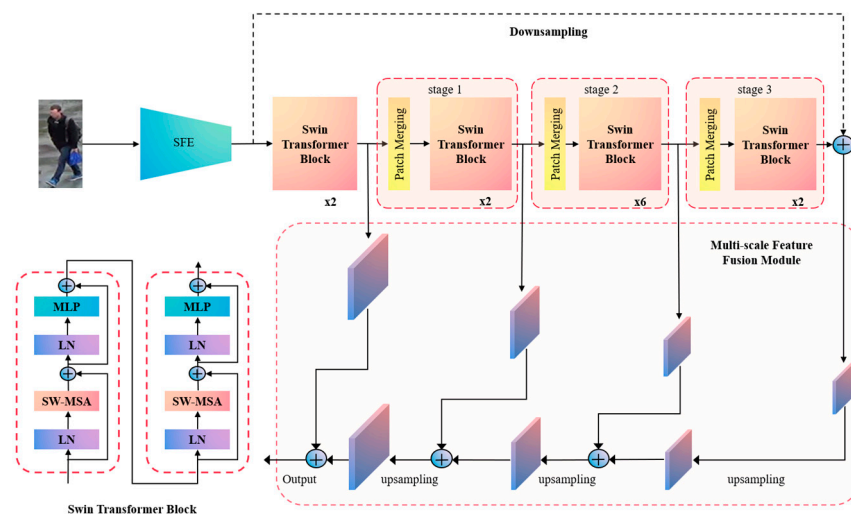


Figure 2. Structure of the multi-scale feature fusion global branch.

- (1) **Shallow Feature Extraction (SFE).** As indicated in [41], convolutional layers excel at image feature extraction, thereby producing a more robust feature representation. The shallow feature extraction module introduced in this paper is depicted in Figure 3.

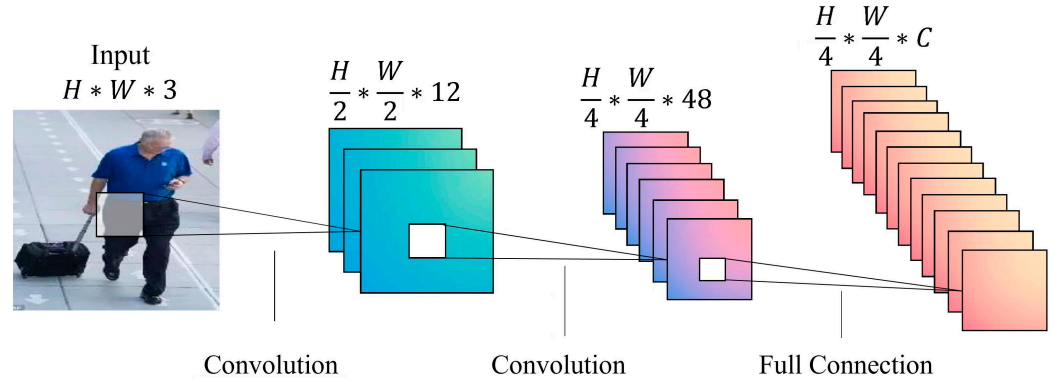


Figure 3. Process of shallow feature extraction module (SFE).

Initially, an image of a pedestrian is denoted as $I \in R^{H \times W \times 3}$ (H represents height, W represents width, and 3 represents the number of channels in the image), is processed through two convolution layers with a convolution kernel size of 2×2 and stride of 2 for the initial feature extraction, resulting in shallow feature expressions. Subsequently, the number of feature channels is mapped to C via a fully connected layer, and it is fed to the Swin Transformer Block [37] for deep feature extraction. Following this, deep and shallow features are fused to strengthen the integration of local and global feature representations. Finally, the combined features are introduced into the MFF to derive a distinctive feature representation.

- (2) **Multi-scale Feature Fusion (MFF).** The backbone network employs a layer-by-layer abstraction approach to extract the target features, progressively expanding the field of view of the feature map through convolution and downsampling algorithms. Incorporating the concept of a feature pyramid [42], we introduce a multi-scale fusion module to enhance feature representations at various levels, capturing both local fine-grained features and global feature information of pedestrians. Additionally, this module is compatible with different scales of the Swin Transformer models.

As depicted in Figure 2, within the MFF module, we independently extract features from the outputs of the different stages in the backbone network. Because the feature maps from the different stages possess distinct sizes, we apply bilinear interpolation to upsample the extracted features sequentially. Subsequently, the original features, features generated by the SFE, and upsampled features are merged using the Add operation. This fusion of multi-level features with distinct perceptual domains enables the capturing of diverse semantic information. Finally, the fused features are fed into a fully connected classifier for the ultimate output.

3.2. Hybrid Attention Local Branch

The hybrid attention local branch is used to cope with the scene of illumination differences and scale changes. We process the original image using random grayscale and random scaling to simulate the illumination difference scene. Building on this, we introduce a parallel hybrid module that combines channel and spatial attention. This module enhances the capability of the model to emphasize crucial features while optimizing computational efficiency.

As depicted in Figure 4, the processed image is first input into OSNet [8] for feature extraction, and then the extracted feature $F \in R^{c \times h \times w}$ is input into the hybrid attention module for maximum pooling and average pooling operations to compress the channels and output two-dimensional maps $F_{s_max} \in R^{1 \times h \times w}$ and $F_{s_avg} \in R^{1 \times h \times w}$. Maximum pooling helps to retain significant information by selecting the largest pixel value in a

specific region to represent the region. Average pooling computes the average value of pixels within a specific region, representing the region effectively. This approach mitigates errors caused by limited area sizes during feature compression, ensuring the preservation of global information within the feature. After that, the generated two-dimensional map is element-wise multiplied with the original feature map to acquire features encompassing both salient and global information within the channel dimension. The calculation process is shown in Equations (1) and (2):

$$F'_{s_max} = \sigma((MaxPool(F))) \otimes F \quad (1)$$

$$F'_{s_avg} = \sigma((AvgPool(F))) \otimes F \quad (2)$$

where $\sigma(\cdot)$ indicates the sigmoid function, and $\otimes$ indicates element-by-element multiplication.

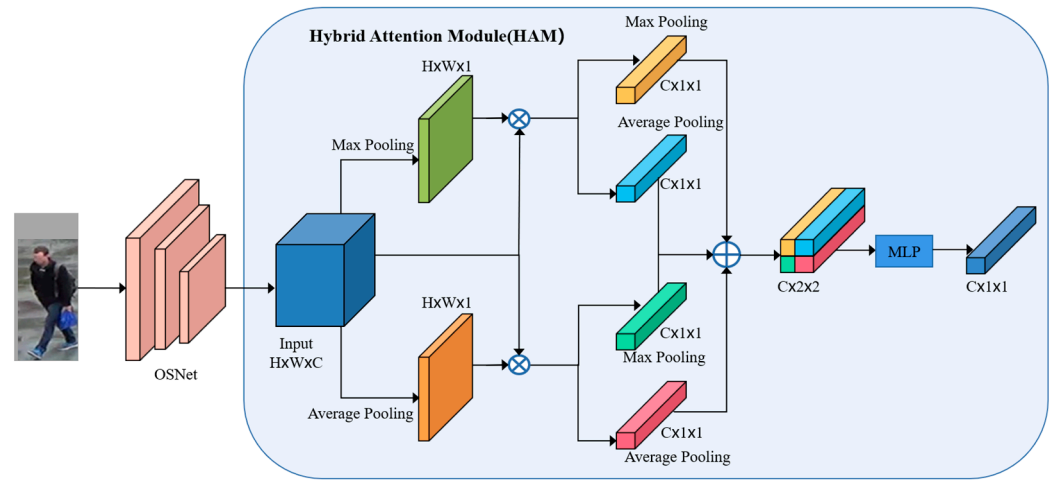


Figure 4. Hybrid Attention Module.

We compress the spatial dimensions of F'_{s_avg} and F'_{s_max} to perform feature learning in the channel dimension, and then obtain four spatial context feature descriptors, merge them, and fuse them through Multi-Layer Perception (MLP) to generate the final attention feature vector w_{sc} . The calculation process is as follows:

$$w_{sc} = \sigma \left(MLP \left(AvgPool \left(F'_{s_avg} \right) \oplus MaxPool \left(F'_{s_avg} \right) \oplus AvgPool \left(F'_{s_max} \right) \oplus MaxPool \left(F'_{s_max} \right) \right) \right) \quad (3)$$

where $\sigma(\cdot)$ indicates the sigmoid function and $\oplus$ indicates concatenation.

3.3. Saliency Guidance Local Branch

The saliency guidance local branch is used to deal with the occlusion in the person re-ID. We apply data augmentation to the input images of this branch to generate a sufficient number of occluded samples for learning. The main idea of the Anti-erasure Federated Block Network (AFBN) proposed in this paper is to utilize two adversarial classifiers driven by feature erasure to find complementary object regions. The architecture of AFBN is depicted in Figure 5.

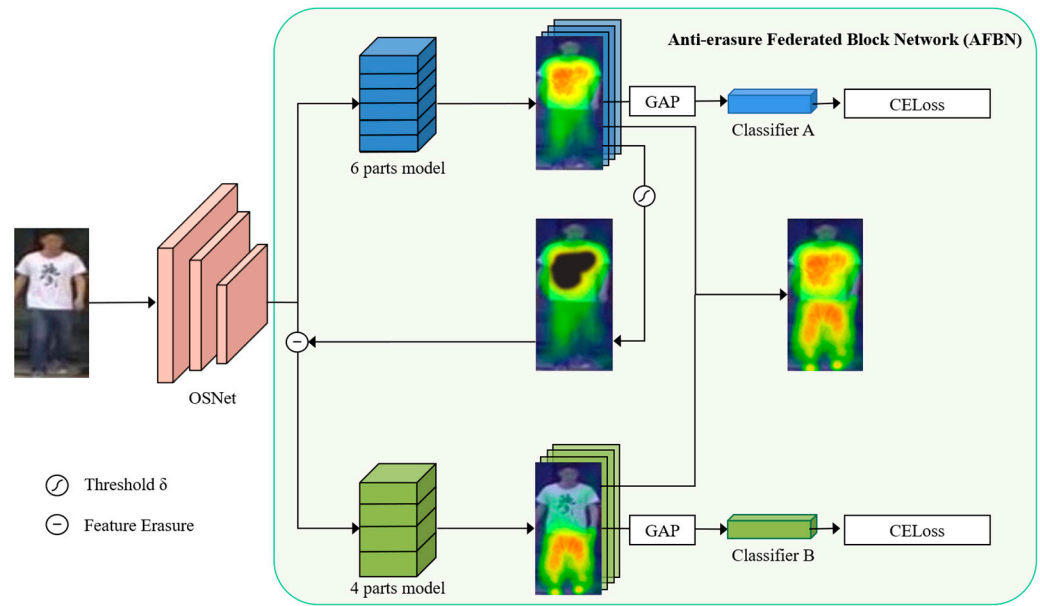


Figure 5. Structure of AFBN.

The AFBN uses the OSNet of [8] as the backbone feature extraction network, and the feature maps extracted from the backbone network are input into parallel classification branches, each consisting of block networks. In order to avoid block overlapping and invalidation of the same region after erasure, the number of blocks in the two branches is different.

We propose Algorithm 1 for AFBN that identifies distinctive regions on the feature map generated by Classifier A by applying a predefined threshold. This identification process guides the zero-value replacement of the input features in Classifier B within the identified regions. Following this, a global average pooling layer and softmax layer are connected for classification. Ultimately, an integrated feature map is obtained by combining the complementary feature maps generated by the two branches.

The algorithm of AFBN is shown as follows.

Algorithm 1: AFBN

Input: Training data $I = \{(I_i, y_i)\}_{i=1}^N$, threshold δ

1: **while** training does not converge **do**

2: $S \leftarrow f(\theta_0, I_i)$ ▷ Extract initial features S_0

3: $M^A \leftarrow f(\theta_A, S, y_i)$ ▷ Extract features M^A

4: $R = M^A > \delta$ ▷ Obtain high discriminative regions R

5: $\bar{S} \leftarrow \text{erase}(S, R)$ ▷ Obtain erased features $\bar{S}_0$

6: $M^B \leftarrow f(\theta_B, \bar{S}, y_i)$ ▷ Extract features M^B

7: $\bar{M}_{fuse} = \max(\bar{M}^A, \bar{M}^B)$ ▷ Obtain fusion feature $\bar{M}_{fuse}$

8: **end while**

Output: $\bar{M}_{fuse}$

$I = \{(I_i, y_i)\}_{i=1}^N$ represents the training image set, where y_i is the label of the image I_i , and N represents the number of images. The input image I_i is first transformed by the backbone network $f(\theta_0)$ into a spatial feature map $S \in R^{H_1 \times W_1 \times K}$ with K channels and $H_1 \times W_1$ resolution. θ represents the learnable parameters of the neural network. $f(\theta_A)$ represents Classifier A, which can generate an object feature map $M^A \in R^{H_2 \times W_2}$ of size $H_2 \times W_2$, and M^A is used to highlight the discriminative feature region of the target.

3.4. Multiple Loss

The cross-entropy loss is used as the loss of the classification model to measure the dissimilarity between the predicted values and the actual ground truth. In the context of the person re-ID task, a smaller cross-entropy value indicates that the model's judgment of the pedestrian category aligns more closely with the true label. The cross-entropy loss function L_{ce} is defined as follows in Equation (4):

$$L_{ce} = \frac{1}{N} \sum_i L_i = -\frac{1}{N} \sum_i \sum_{c=1}^M y_{ic} \log(p_{ic}) \quad (4)$$

where M indicates the number of categories (person ID), y_{ic} indicates 1 when the true category of sample i is c , otherwise 0, and p_{ic} represents the probability that the sample i is predicted as class c .

Hard sample triplet loss is used as the metric loss. In every training batch, P pedestrians are randomly chosen, and K images are selected for each pedestrian. Therefore, a batch contains $P \times K$ pictures in total. For each image a in the batch, a hard positive sample and a hard negative sample are selected to create a triplet. Here, the set of images from the same pedestrian as a is denoted as A , and the set of images from different pedestrians as a is denoted as B . The hard sample triplet loss function $L_{TriHard}$ is defined as follows in Equation (5):

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

where a is the threshold parameter set independently. In contrast to the traditional triplet loss, the hard sample triplet loss can expedite the convergence of the network and facilitate the learning of superior representations.

KL divergence is used to strengthen the information exchange between different branches and avoid model collapse. Using KL divergence allows each branch to propagate its own knowledge to other branches. By learning from each other, the local branches can gain the ability to perceive specific scenes. Meanwhile, the global branch gathers insights from a broader range of scenes, thereby decreasing the likelihood of overfitting and enhancing the overall robustness. Equation (6) is used to calculate the probability p that the category of the pedestrian sample x_i is m .

$$p_m(x_i) = \frac{\exp(\gamma W_m f)}{\sum_{k=1}^M \exp(\gamma W_k f)} \quad (6)$$

where M denotes the total number of pedestrian IDs, γ denotes the scale factor, which is kept constant to expedite the training process, W_m and W_k are the weight vectors related to categories m and k in the final classification layer, and f denotes the normalized features. Therefore, the KL divergence of the current branch can be expressed as shown in Equation (7):

$$L_{ML}^{\Theta} = \frac{1}{S-1} \sum_{s=1, s \neq \Theta}^S \sum_{i=1}^N \sum_{m=1}^M p_m(x_i) \log \frac{p_m^{\Theta}(x_i)}{p_m^s(x_i)} \quad (7)$$

where Θ indicates the current branch, and S represents the total number of branches.

The final loss function for each branch is formulated as follows:

$$L^{\Theta} = L_{ML}^{\Theta} + L_{ce} + L_{TriHard} \quad (8)$$

4. Experiments

4.1. Datasets

Our model is evaluated on three benchmark datasets which are created for the person Re-ID task: Market1501 [43], DukeMTMC-reID [44], and MSMT17 [45].

Market1501: Market1501 collects pedestrian images captured by 6 cameras, with a total of 1,501 individuals annotated, resulting in 32,668 images. Specifically, 12,936 images are

annotated for 751 pedestrians in the training set, whereas the test set included 19,732 images annotated for 750 pedestrians. Notably, there are no overlapping pedestrian IDs between the training and test sets, meaning none of the 751 pedestrians present in the training set is duplicated in the test set.

DukeMTMC-reID: DukeMTMC-reID is a subset of images captured every 120 frames from the videos of the DukeMTMC [46] dataset. In total, the DukeMTMC-reID dataset comprise 36,411 images depicting 1812 pedestrians. Within this dataset, 1404 pedestrians are captured by more than 2 cameras, whereas 408 pedestrians are exclusively captured by a single camera.

MSMT17: MSMT17 is a large dataset similar to real scenes. Using 15 cameras, including 12 outdoor cameras and 3 indoor cameras, approximately 126,441 images are collected under different lighting conditions and background interference, including 4101 different pedestrian identities, and each pedestrian has multiple image samples.

4.2. Experimental Parameters

Our MFFNet method is trained end-to-end with a weight decay of 5×10^{-4} and a batch size of 16 for 200 epochs. The learning rate is set to 0.05 for the three benchmark datasets. We resize the input frame to 224×224 pixels. In order to simulate the scenes with lighting differences, a random grayscale is used to process the original image in the partial branch of the mixed attention area, with the probability set to 0.3.

To generate images with scale variations, random scaling is performed on the duplicates of the original input image. The image is scaled to sizes ranging from 0.8 to 1.1 times the original size, and the scaling factors are randomly generated. The model is implemented using PyTorch.

4.3. Experimental Results

Table 1 shows the Rank-1 and mAP comparisons of our MFFNet on the Market1501 [43], DukeMTMC-reID [44], and MSMT17 [45] datasets. Our approach yields a better performance than several other SOTA methods on these three datasets.

Table 1. Comparison of Rank-1 and mAP performance between our approach and several state-of-the-art methods on the Market1501, DukeMTMC-reID, and MSMT17 datasets.

Method	Venue	Market-1501		DukeMTMC-reID		MSMT17	
		Rank-1/%	mAP/%	Rank-1/%	mAP/%	Rank-1/%	mAP/%
SVDNet [31]	ICCV 17	82.3	62.1	76.7	56.8	-	-
PCB [32]	ECCV18	92.3	77.4	81.7	66.1	-	-
PCB+RPP [32]	ECCV18	93.8	81.6	83.3	69.2	-	-
MHN [26]	CVPR19	95.1	85.0	89.1	77.2	-	-
Auto-ReID [30]	ICCV19	94.5	85.1	88.5	75.1	78.2	52.5
OSNet [8]	ICCV19	94.8	84.9	88.6	73.5	78.7	52.9
HOReID [33]	CVPR20	94.2	84.9	86.9	75.6	78.4	54.8
SNR [28]	CVPR20	94.4	84.7	84.4	72.9	-	-
CDNet [29]	CVPR21	95.1	86.0	88.6	76.8	78.9	54.7
L3DS [27]	CVPR21	95.0	87.3	88.2	76.1	-	-
FED [25]	CVPR22	95.0	86.3	89.4	78.0	-	-
IER+MTOR [22]	IEEE22	90.2	75.5	82.6	66.4	-	-
FPTE [34]	Springer23	94.3	87.0	88.7	77.2	-	-
MFFNet	This work	95.2	85.8	92.5	80.6	80.2	56.6

The best performances in each dataset are highlighted in red, whereas the second-best is highlighted in blue.

On Market1501, we achieve 95.2% and 85.3% on Rank-1 and mAP, respectively, and Rank-1 performs better than the other methods. For the DukeMTMC-reID, we achieve 92.5% and 80.6% on Rank-1 and mAP, respectively, which are significantly better than the performance of other methods. Additionally, on the MSMT17, which is closest to a real scene, MFFNet also exerts its structural advantages. Compared with CDNet, which has the

best performance of Rank-1, and the HOREID, which has the best performance of mAP, it still increases by 1.3% and 1.8%, respectively. Overall, our model yields better performance, thus demonstrating the superiority and effectiveness of our approach.

In order to compare the performance of the model more intuitively, we visualize the recognition results on the Market-1501 dataset and compare them with the results from other methods. These visualizations are presented in Figures 6–8. Notably, MFFNet demonstrates excellent recognition efficacy and robustness, particularly under illumination differences, scale changes, and occlusion scenes. In each row, the first column of images represents the target pedestrian for the query, while the subsequent ten columns depict the recognition outcomes sorted based on similarity. The green solid line indicates that the pedestrian in the image corresponds to the queried pedestrian ID, whereas the red dashed line indicates a different pedestrian ID from that of the query.

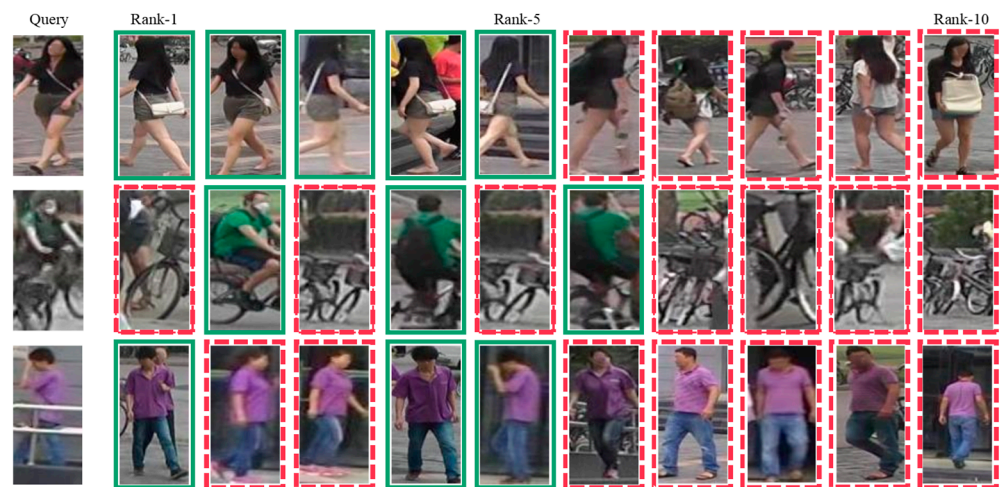


Figure 6. Visualization results of PCB on Market-1501.



Figure 7. Visualization results of HOREID on Market-1501.

Additionally, we employ Grad-CAM [47] to generate heat maps to visualize the features extracted by each model. The visualization results are presented in Table 2, where the color gradient from blue to red indicates the feature weight values ranging from the lowest to the highest. MFFNet can focus on pedestrian targets more completely and accurately, basically covering the entire pedestrian area. At the same time, it has better anti-interference ability in complex scenes.

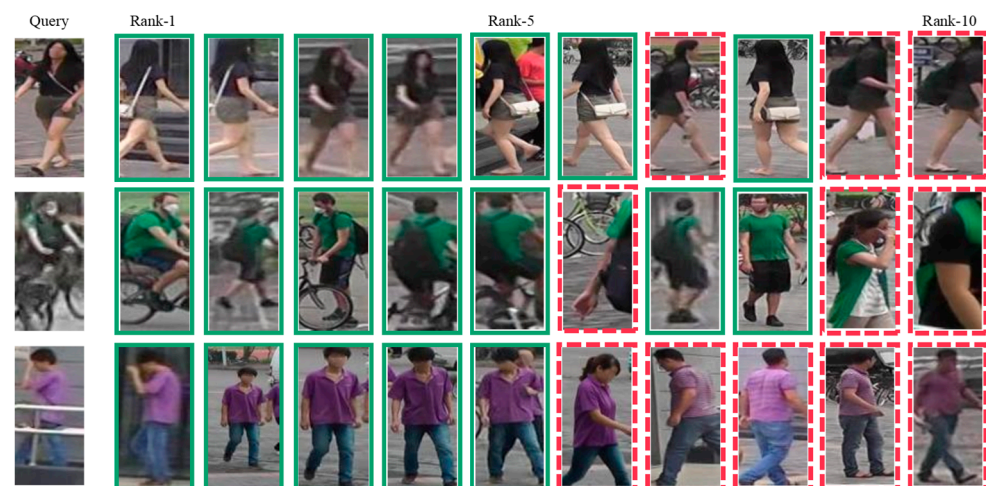


Figure 8. Visualization results of MFFNet on Market-1501.

Table 2. Heat maps of visualization results of feature extraction based on Grad-CAM.

Model	Standard Scene	Occlusion Scene	The Scene with Illumination Difference and Scale Change
Input			
ResNet50 [48]			
Swin Transformer [37]			
MFFNet			

4.4. Ablation Studies

Effects of different branches in the model. We conduct experiments on the effects of each branch of the model on the MSMT17 dataset. As depicted in Table 3, the combination of different branches significantly improves the accuracy of recognition. In the two-branch network, the saliency guidance local branch substantially boosts model performance. Notably, the MFFNet model incorporating all the three branches achieve the highest performance. This highlights the diverse features extracted by the two local branches and validates the effectiveness of each branch.

Table 3. Effects of different branches in the MFFNet model.

Branch	Rank-1/%	mAP/%
Multi-scale Fusion global branch	78.3	52.7
Multi-scale Fusion global branch + Hybrid Attention local branch	79.1	53.8
Multi-scale Fusion global branch + Saliency Guidance local branch	79.5	55.3
Multi-scale Fusion global branch + Hybrid Attention local branch + Saliency Guidance local branch	80.2	56.6

Effects of different modules in branches. To validate the efficacy of each module within each branch, we conduct experiments using the MSMT17 dataset. Table 4 illustrates the impact of the shallow feature extraction module (SFE) and multi-scale feature fusion module (MFF) on the performance of the global branch, confirming the effectiveness of the multi-level feature fusion strategy.

Table 4. Effect of different modules on the global branch.

Method	Rank-1/%	mAP/%
Swin Transformer	77.0	51.5
Swin Transformer + SFE	77.2	51.8
Swin Transformer + MFF	78.1	52.3
Swin Transformer + SFE + MFF	78.3	52.7

As depicted in Table 5, we conduct a comparison between the improved Hybrid Attention Module (HAM) and the standard Fusion Attention Module, as well as the baseline model without any attention module, within the Hybrid Attention Local Branch. This comparison validates the effectiveness of the proposed method.

Table 5. Effect of different modules in the hybrid attention branch.

Method	Rank-1/%	mAP/%
OSNet	78.3	52.8
OSNet + Standard Fusion Attention Module	78.9	53.6
OSNet + HAM	79.1	53.8

In the saliency guidance local branch, we employ a combination of adversarial erasing and partitioned networks to extract secondary salient and local features, thereby improving the recognition ability in occlusion scenarios. As evidenced in Table 6, the experimental results demonstrate a significant performance improvement achieved through the utilization of the AFBN introduced in this paper.

Table 6. Effects of different modules in the saliency guidance branch.

Method	Rank-1/%	mAP/%
OSNet	78.3	52.8
Random Erasing	78.5	52.8
Adversarial Erasing	78.8	53.9
Random Erasing + partitioned network	79.0	54.3
Adversarial Erasing + partitioned network (AFBN)	79.5	55.3

4.5. Discussions

To address the problem of person re-ID in complex scenarios, we employ three different branches to capture multi-scale features and simulate real-world challenges, such as illumination differences, scale changes, and occlusions. Through experiments, the feasibility and effectiveness of the MFFNet are demonstrated. The method proposed by

Liao et al. [7] utilizes saliency model-based multi-scale feature fusion to extract significant body regions from images to mitigate interference from complex backgrounds in person re-ID tasks. Wang et al. [25] presented a Feature Erasing and Diffusion Network capable of addressing both non-target pedestrian occlusions and non-target pedestrians. Zhou et al. [34] combined the complementary advantages of CNN and Transformer, using feature enhancement and pose-guided modules to tackle challenges posed by poses and occlusions in person Re-ID. Compared to these methods, our approach takes into account various complex factors that influence a person re-ID. In terms of the experimental results, on DukeMTMC-reID, MFFNet outperformed recent methods in both Rank-1 and mAP. On Market1501, Rank-1 of MFFNet is 0.5% lower than [25] and 1.2% lower than [34]. Comparisons of specific experimental results are shown in Figures 9 and 10. This indicates that MFFNet demonstrates superior recognition accuracy in typical scenarios, accurately matching the correct identities of pedestrians. However, in simple scenarios, it falls slightly behind [25,34] in the overall evaluation of all query results. Furthermore, we conduct experiments on MSMT17, which is more challenging and closer to a real scene, and the results show significant improvements, as shown in Table 1. In future work, it will be necessary to conduct research on recognizing pedestrians in more scenarios and different categories, further enhancing the performance of the model.

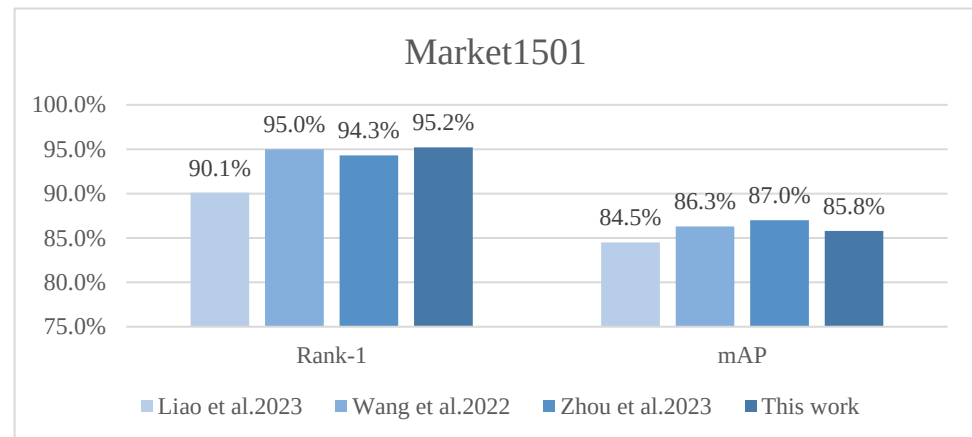


Figure 9. Comparison results between the models in [7,25,34] and our model on Market1501.

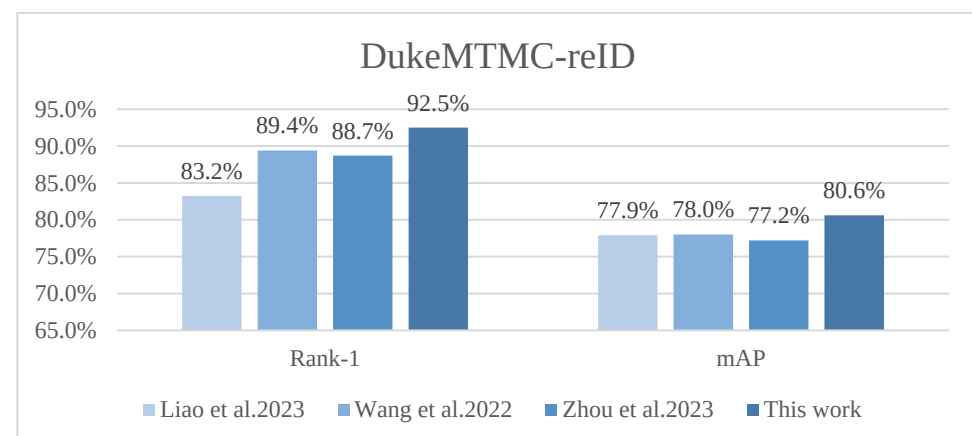


Figure 10. Comparison results between the models in [7,25,34] and our model on DukeMTMC-reID.

5. Conclusions and Prospect

In this paper, we propose a person re-ID method based on a multi-branch feature fusion, MFFNet, comprising three distinct branches to address the problems of person re-ID in different and complex scenarios. Our MFFNet aggregates shallow and deep features through a multi-scale feature fusion module and enhances its recognition capabilities using

a combination of hybrid attention and saliency guidance mechanisms. Additionally, we incorporate KL divergence to enhance the information exchange between different branches and improve the robustness of the features. Extensive experimental evaluations of standard benchmarks demonstrate that our model outperforms the existing models.

We only consider the impact of typical complex scenes, such as scale changes, illumination differences, and occlusions, on person re-ID. Therefore, in future work, we plan to optimize the model to adapt it to other complex scenarios without excessively expanding the model. In addition, employing different training strategies, such as unsupervised or self-supervised learning methods, to enhance the applicability of this model in real-world scenarios is also a focus of our future work.

Author Contributions: Conceptualization: P.L. and R.T.; methodology: X.H. and X.W.; resources: R.T. and P.L.; data curation: X.H.; writing—original draft preparation: X.W.; writing—review and editing: P.L. and X.H.; and supervision: P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the National Key R&D Program of China (No.2022YFB3305800-3). The project is called “Development and Demonstration of a Comprehensive Control Platform for Product Recycling, Disassembly, and Reutilization”.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created. Data are available in a publicly accessible repository. The data presented in this study are openly available in [43–45].

Acknowledgments: We are grateful to the anonymous reviewers for their comments on this manuscript.

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

References

1. 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.
2. 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, Seattle, WA, USA, 27–30 June 2016.
3. Wu, L.; Shen, C.; Hengel, A.v.d. Personnet: Person re-identification with deep convolutional neural networks. *arXiv* **2016**, arXiv:1601.07255.
4. Fu, Y.; Wei, Y.; Zhou, Y.; Shi, H.; Huang, G.; Wang, X.; Yao, Z.; Huang, T. Horizontal pyramid matching for person re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019.
5. Miao, J.; Wu, Y.; Liu, P.; Ding, Y.; Yang, Y. Pose-guided feature alignment for occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
6. Chen, Y.; Zhu, X.; Gong, S. Person re-identification by deep learning multi-scale representations. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017.
7. Liao, K.; Wang, K.; Zheng, Y.; Lin, G.; Cao, C. Multi-scale saliency features fusion model for person re-identification. *Multimed. Tools Appl.* **2023**, 1–16. [[CrossRef](#)]
8. 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.
9. 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/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
10. Jia, M.; Sun, Y.; Zhai, Y.; Cheng, X.; Yang, Y.; Li, Y. Semi-attention Partition for Occluded Person Re-identification. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023.
11. Ning, X.; Gong, K.; Li, W.; Zhang, L. JWSAA: Joint weak saliency and attention aware for person re-identification. *Neurocomputing* **2021**, 453, 801–811. [[CrossRef](#)]
12. Zhang, Z.; Zhang, H.; Liu, S. Person re-identification using heterogeneous local graph attention networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
13. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.

14. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Saliency-guided cascaded suppression network for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020.
15. Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; Tian, Q. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [[CrossRef](#)] [[PubMed](#)]
16. Zhang, G.; Chen, C.; Chen, Y.; Zhang, H.; Zheng, Y. Transformer-based global-local feature learning model for occluded person re-identification. *J. Vis. Commun. Image Represent.* **2023**, *95*, 103898. [[CrossRef](#)]
17. Wang, G.; Yuan, Y.; Chen, X.; Li, J.; Zhou, X. Learning discriminative features with multiple granularities for person re-identification. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018.
18. Ming, Z.; Yang, Y.; Wei, X.; Yan, J.; Wang, X.; Wang, F.; Zhu, M. Global-local dynamic feature alignment network for person re-identification. *arXiv* **2021**, arXiv:2109.05759.
19. Li, X.; Zheng, W.-S.; Wang, X.; Xiang, T.; Gong, S. Multi-scale learning for low-resolution person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 11–18 December 2015.
20. 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, 18–23 June 2018.
21. 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.
22. Zhang, G.; Luo, Z.; Chen, Y.; Zheng, Y.; Lin, W. Illumination unification for person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 6766–6777. [[CrossRef](#)]
23. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Vrstc: Occlusion-free video person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.
24. Huang, H.; Li, D.; Zhang, Z.; Chen, X.; Huang, K. Adversarially occluded samples 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.
25. 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.
26. Chen, B.; Deng, W.; Hu, J. Mixed high-order attention network for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
27. Chen, J.; Jiang, X.; Wang, F.; Zhang, J.; Zheng, F.; Sun, X.; Zheng, W.-S. Learning 3D shape feature for texture-insensitive person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
28. Jin, X.; Lan, C.; Zeng, W.; Chen, Z.; Zhang, L. Style normalization and restitution for generalizable person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 14–19 June 2020.
29. Li, H.; Wu, G.; Zheng, W.-S. Combined depth space based architecture search for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021.
30. Quan, R.; Dong, X.; Wu, Y.; Zhu, L.; Yang, Y. Auto-reid: Searching for a part-aware convnet for person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019.
31. 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.
32. 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.
33. Wang, P.; Zhao, Z.; Su, F.; Zu, X.; Boulgouris, N.V. HOREID: Deep high-order mapping enhances pose alignment for person re-identification. *IEEE Trans. Image Process.* **2021**, *30*, 2908–2922. [[CrossRef](#)] [[PubMed](#)]
34. Zhou, S.; Zou, W. Fusion pose guidance and transformer feature enhancement for person re-identification. *Multimed. Tools Appl.* **2023**, 1–19. [[CrossRef](#)]
35. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
36. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021.
37. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Virtual, 11–17 October 2021.
38. 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.
39. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019.

40. Woo, S.; Park, J.; Lee, J.-Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
41. Xiao, T.; Singh, M.; Mintun, E.; Darrell, T.; Dollár, P.; Girshick, R. Early convolutions help transformers see better. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 30392–30400.
42. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017.
43. Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Bu, J.; Tian, Q. Person re-identification meets image search. *arXiv* **2015**, arXiv:1502.02171.
44. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
45. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
46. 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.
47. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017.
48. 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, Seattle, WA, USA, 26–27 June 2016.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.