

Article

MIX-Net: Hybrid Attention/Diversity Network for Person Re-Identification

Minglang Li ¹, Zhiyong Tao ^{1,*}, Sen Lin ² and Kaihao Feng ¹

¹ School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China; 472220498@stu.lntu.edu.cn (M.L.); 472220492@stu.lntu.edu.cn (K.F.)

² School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang 110158, China; linsen@sylu.edu.cn

* Correspondence: taozhiyong@lntu.edu.cn

Abstract: Person re-identification (Re-ID) networks are often affected by factors such as pose variations, changes in viewpoint, and occlusion, leading to the extraction of features that encompass a considerable amount of irrelevant information. However, most research has struggled to address the challenge of simultaneously endowing features with both attentive and diversified information. To concurrently extract attentive yet diverse pedestrian features, we amalgamated the strengths of convolutional neural network (CNN) attention and self-attention. By integrating the extracted latent features, we introduced a Hybrid Attention/Diversity Network (MIX-Net), which adeptly captures attentive but diverse information from personal images via a fusion of attention branches and attention suppression branches. Additionally, to extract latent information from secondary important regions to enrich the diversity of features, we designed a novel Discriminative Part Mask (DPM). Experimental results establish the robust competitiveness of our approach, particularly in effectively distinguishing individuals with similar attributes.

Keywords: person Re-ID; deep learning; feature-dropping; hybrid attention



Citation: Li, M.; Tao, Z.; Lin, S.; Feng, K. MIX-Net: Hybrid Attention/Diversity Network for Person Re-Identification. *Electronics* **2024**, *13*, 1001. <https://doi.org/10.3390/electronics13051001>

Academic Editor: Daniele Riboni

Received: 10 February 2024

Revised: 3 March 2024

Accepted: 4 March 2024

Published: 6 March 2024



Copyright: © 2024 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 (Re-ID) is a typical image retrieval problem that aims to locate the same target across different cameras and locations using various technologies such as computer vision, pattern recognition, and deep learning. Recently, person re-identification has garnered significant attention in research circles due to its forensic and commercial relevance. Person Re-ID technology has broad applications in areas such as video surveillance and criminal investigation, compensating for the limitations of inherent recognition devices for faces or other biometric features. A common approach to person Re-ID is to directly extract features from the entire person [1], followed by fine-tuning on a specific person Re-ID network and dataset using deep neural networks pre-trained on ImageNet. Despite extensive research in recent years, challenges persist due to pose variations, background clutter, and occlusion issues [2]. Extracting the most discriminative features and designing effective feature-matching algorithms are critical for addressing person Re-ID tasks.

Before 2016, feature extraction for person Re-ID mainly involved extracting low-level visual features, including HOG features, color histograms, and key points. However, these algorithms for extracting low-level visual features encounter difficulties in capturing highly discriminative features when confronted with diverse image samples [3]. After 2016, with the development of deep learning, person Re-ID technology has made significant breakthroughs, resulting in substantial improvements in recognition accuracy. With respect to feature extraction, convolutional neural networks (CNNs) integrated with attention mechanisms can identify salient regions within images [4], mitigating the subjectivity issues associated with traditional methods. In terms of models, deep learning-trained

models can explore deeper levels of information, learning associations between samples more profoundly.

Compared to single-stream networks, feature fusion networks that incorporate features of different scales and weights exhibit superior performance. Nevertheless, these feature fusion networks typically encounter a dilemma between focusing on extracting information from specific salient regions and achieving diverse information extraction across the entire global context. Emphasizing information extraction from specific salient regions enhances the accuracy of identifying individuals with particular attributes but may neglect global information. Conversely, choosing diverse information extraction across the global context ensures the inclusion of information from most regions, but the resulting features may contain excessive irrelevant information, potentially impeding the final decision-making process. Some researchers argue that employing a quantum-based framework in certain computer vision tasks can balance network efficiency and performance [5]. However, in relatively complex person re-identification tasks, it is essential to enhance the salient information in pedestrian images while suppressing irrelevant details to improve the discriminative power of the final features. Therefore, attention mechanisms have been proven to be a viable solution. Inspired by suppression networks and attention mechanisms [6–8], this paper introduces a Discriminative Part Mask (DPM) designed to mine latent information, which is combined with the Mix branch of the hybrid attention module. The proposed person re-identification network, named MIX-Net, has been experimentally validated on five large datasets and compared with state-of-the-art person re-identification methods.

The research strategy of this paper is as follows: Part 1 provides an introduction to this study. Section 2 reviews related work, including person Re-ID methods and datasets. Section 3 provides a comprehensive description of the architecture and features of MIX-Net. Section 4 details comparative experiments and ablation studies against other methods to evaluate the performance of our network. Section 5 concludes the paper.

The contributions of this paper can be summarized as follows:

- The design of MIX-Net, a person Re-ID network capable of extracting diverse information by combining hybrid attention branches and latent information extraction branches;
- The proposal of a Discriminative Part Mask (DPM) attention suppression module to extract discriminative information from non-salient regions;
- Extensive experiments on Market-1501, DukeMTMC, MSMT17, Occluded-Duke, and Occluded-REID datasets. MIX-Net outperforms existing methods, achieving competitive results in all popular benchmarks. Through rigorous ablation studies and visualization, we validate the helpfulness of each module and branch for our task.

2. Relate Works

2.1. Deep Learning-Based Person Re-Identification

Person re-identification (Re-ID) aims to locate the same target across different cameras and locations, finding applications in video surveillance and criminal investigations. A common approach in Re-ID involves extracting features directly from the entire person, leveraging deep neural networks pre-trained on ImageNet, and fine-tuning specific Re-ID networks and datasets. Despite extensive research in recent years, Re-ID remains an ongoing challenge due to factors such as pose variations, background clutter, and occlusion [9].

To address these challenges, various researchers have proposed diverse solutions. Liao et al. [10] introduced a method known as Local Maximal Occurrence (LOMO). This approach incorporates both color histograms and SILTP histograms, representing pedestrian appearance features through a combination of maximum pooling, scale operations, and logarithmic transformations. Koestinger et al. [11] extracted image color histograms and texture features, concatenating various features to represent images. They employed PCA for dimensionality reduction to obtain a low-dimensional representation. Additionally, they proposed a regularized smoothing KISS metric method to achieve image recognition in the

low-dimensional space. Yi et al. [12] were the first to apply Siamese Convolutional Neural Networks (SCNNs) to the field of person re-identification. Considering the variations in background and lighting conditions in person re-identification image data, they abandoned the practice of sharing weights in the original network, making the two sub-networks independent of each other. Ahmed et al. [13] proposed a deep network for person re-identification based on SCNNs. This network takes image pairs as input, calculates the differences in feature maps, and ultimately determines whether the image pairs belong to the same category. Li et al. [14] introduced a Filter Pairing Neural Network (FPNN), which uniformly divides pedestrian images into multiple grids and matches different parts of the same strip to determine the consistency between two pedestrians.

In the initial stages of our research, the outcomes were unsatisfactory. Nevertheless, the advent of deep learning studies has brought about a breakthrough in person re-identification. Wang et al. [15] introduced a convolutional structure called Wconv, where each input image undergoes two independent convolutional layers, resulting in two separate feature maps that are subsequently fused. This process yields distinctive feature maps for each input image. Zhao et al. [16] proposed an approach that eliminates fixed region segmentation methods. Inspired by attention mechanisms, the network is partitioned into K branches based on different weights to extract features from distinct regions, addressing the issue of misaligned key points. Wang et al. [17] proposed an effective method for person re-identification that addresses occlusion issues. Deng et al. [18] combined CycleGAN with a Siamese network, proposing a method for image transfer between datasets. During the process of migrating images from the source dataset to the target dataset, the label information of the images is preserved. Liu et al. [19] introduced UnityGAN, which learns background style differences among different cameras and generates average-style images based on these differences, thereby enhancing the generalization ability of person re-identification models to camera background styles. Su et al. [20] proposed a Pose Feature-Driven Convolutional (PDC) person re-identification model, which consists of two branches and two subnetworks. The two branches are dedicated to learning global feature representations describing the entire image and region-specific feature representations highlighting key local areas. Zhao et al. [16] proposed an efficient lightweight pedestrian alignment network, integrating the use of Fully Convolutional Network (FCN) and Region Proposal Network. This approach extracts the k most discriminative regions in pedestrian images and forms the final pedestrian feature representation through concatenation. Liu et al. [21] devised a pose transfer method by selecting a dataset with diverse poses, utilizing a pose-detection algorithm to extract pedestrian pose information represented using RGB images, and combining this with pedestrian images from the target dataset as input data to train a Conditional Generative Adversarial Network (CGAN) [22] to generate new pedestrian images, thereby achieving pose information transfer. Tian et al. [23] proposed the Variational Self-Distillation (VSD) model, which effectively filters out redundant information such as background while extracting discriminative features.

Attention mechanisms [24] effectively eliminate background interference, enabling the network to concentrate on crucial regions of personal images. Strategies incorporating both local and global features enhance network robustness. Song et al. [25] utilized a pre-trained binary mask as an additional channel in the model input, forming a four-channel input model named RGB-Mask, alongside RGB. The mask divides the complete image into the background and pedestrian body regions. The network is supervised with triplet loss to learn features from the pedestrian region while disregarding background features. Li et al. [26] introduced a spatiotemporal attention model incorporating multiple spatial attention models and diverse regularization terms to ensure the learning of various body parts. Building upon this, a temporal attention model is employed to fuse image features across the sequence, effectively addressing challenges such as pedestrian occlusion and misalignment in video sequences. Zhao et al. [16] employed a spatial Transformer network as a hard attention mechanism for efficient region searching. Li et al. [27] introduced a multi-scale attention selection mechanism to address issues like poor pixel-level boundary localization

and noise interference. Song et al. [25] proposed a method using spatial attention maps to generate body awareness and background awareness, removing background clutter. Wang et al. [28] utilized multi-task learning to obtain a series of 3D masks, reweighting feature maps spatially and channel-wise for channel attention allocation. Chen et al. [29] employed a channel attention module to calculate correlation coefficients between different channels, implementing channel-level attention mechanisms. Zhuo et al. [30] introduced the Attention Framework of Person Body (AFPB), which focuses on occluded regions by comparing different types of occluded pedestrians. Addressing dynamic occlusion in video sequences is advantageous compared to handling this issue in image datasets. This is because occluded regions vary across frames in different time sequences, allowing the utilization of different video frames to maximize the completion of pedestrian image information. Liu et al. [31] suggested an innovative convolutional network, termed A-GANet, which incorporates graph relation mining for enhanced performance. They employed an adversarial learning module utilizing a modality discriminator and feature transformer to facilitate cross-modal feature space alignment between text and vision.

2.2. Datasets

Market-1501 [32], introduced by Tsinghua University in 2015, consists of 32,668 annotated images featuring 1501 unique pedestrians recorded by six cameras. Within this dataset, 12,936 images of 751 pedestrians are utilized for training, leaving the remaining images for testing purposes. The test set comprises 3368 images, each representing one of the 750 pedestrians.

DukeMTMC [33], introduced by Duke University in 2016, comprises 36,411 images featuring 1812 pedestrians observed through eight cameras. Among these pedestrians, 1404 are detected in two or more cameras, whereas 408 individuals (considered interfering) are exclusively visible in a single camera. The 1404 pedestrians are randomly divided, with 702 assigned for training and the remainder for testing.

MSMT17 [34], proposed by Peking University in 2018, stands as the most intricate person re-identification dataset to date. It includes 126,441 images of 4101 pedestrians captured by 15 cameras, depicting various weather conditions and time periods.

Occluded-Duke [35] is the largest occlusion dataset to date. The training set comprises 702 individuals with a total of 15,618 images, whereas the query set consists of 519 individuals with 2210 images, and the gallery set includes 1110 individuals with 17,661 images. This dataset represents the most complex occlusion ReID dataset to date, featuring variations in viewpoints and multiple occluding objects, such as cars, bicycles, trees, and other individuals.

Occluded-REID [35], comprising 2000 images captured by a mobile camera, features 200 occluded pedestrians. Each identity (ID) consists of five full-body images and five occluded images. The pedestrian images have been resized to dimensions of 128×64 for uniformity.

3. Materials and Methods

This paper proposes a person re-identification network named MIX-Net based on a hybrid attention mechanism, as illustrated in Figure 1. We choose a modified RESNET-50 as the backbone network for foundational feature extraction from person images. The extracted feature maps undergo training through both a global branch and a hybrid attention branch. To further enhance the network's robustness and capture detailed feature information, we introduce Spectral Value Difference Orthogonality (SVDO) and apply it to both activations and weights [36]. The regularization term for orthogonality in the feature space (O.F.) is designed to decrease feature correlations, providing direct benefits for matching purposes. The orthogonality regularization term on weights (O.W.) fosters diversity in convolutional kernels, thereby augmenting learning capacity.

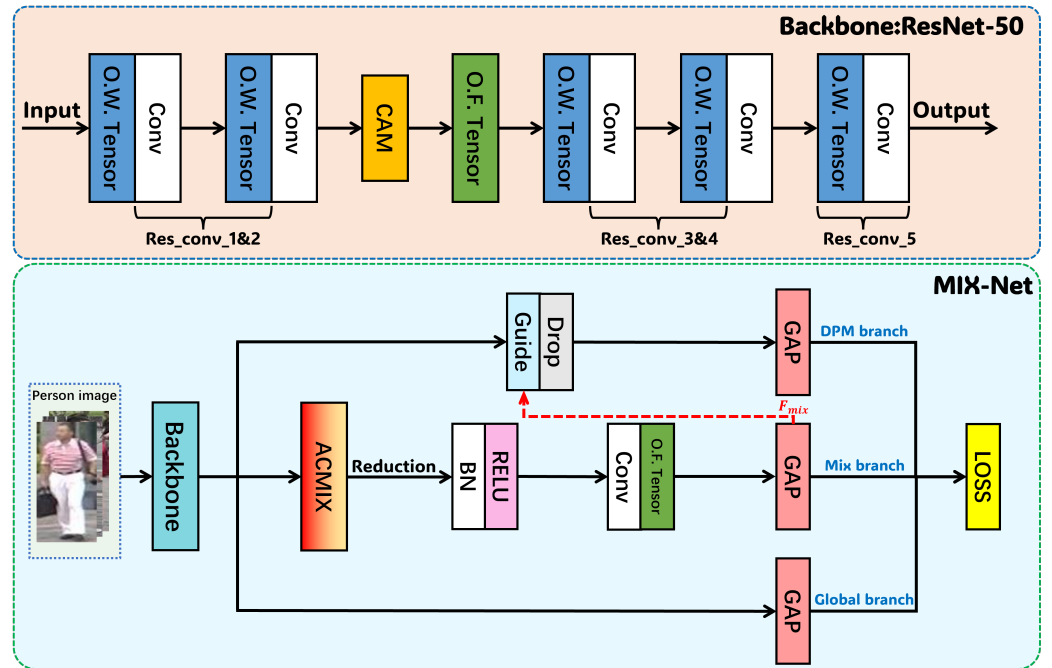


Figure 1. MIX-Net structure diagram. We utilize a RESNET-50 backbone enhanced by CAM, O.F., and O.W. as the primary network. The network is divided into three branches: the global branch, the Mix branch, and the DPM branch. The red dashed arrow indicates that the DPM branch is guided by the output features of the Mix branch.

3.1. CAM

In trained Convolutional Neural Network (CNN) classifiers, high-level convolutional channels are widely acknowledged for their semantic relevance, often exhibiting category selectivity. In the context of person re-identification tasks, these advanced channels display a “grouping” effect, where specific channels exhibit analogous semantic contexts, such as foreground figures, occlusions, or backgrounds, leading to stronger correlations among them. To address this issue, we introduce the Channel Attention Module (CAM), designed to group and aggregate these semantically similar channels [37]. CAM’s primary objective is to emphasize the interdependence of high-level channels, facilitating a better capture of semantic information in personal images. This method of channel grouping and aggregation contributes to the enhancement of the performance of person re-identification models, particularly when dealing with complex scenarios and occlusions. The incorporation of CAM allows the model to focus more precisely on crucial semantic features, thereby improving the accuracy and robustness of person re-identification tasks.

The specific structure of CAM is depicted in Figure 2. For an input feature map $A \in \mathbb{R}^{C \times H \times W}$, where C denotes the total number of channels, and $H \times W$ represents the size of the feature map, the computation of the channel affinity matrix $X \in \mathbb{R}^{C \times C}$ is expressed as follows:

$$x_{ij} = \frac{\exp(A_i \cdot A_j)}{\sum_{j=1}^C \exp(A_i \cdot A_j)}, i, j \in \{1, \dots, C\} \quad (1)$$

where x_{ij} represents the influence of channel i on channel j , the final output feature map E is calculated as follows:

$$E_i = \sum_{j=1}^C (x_{ij} A_j) + A_i, i \in \{1, \dots, C\} \quad (2)$$

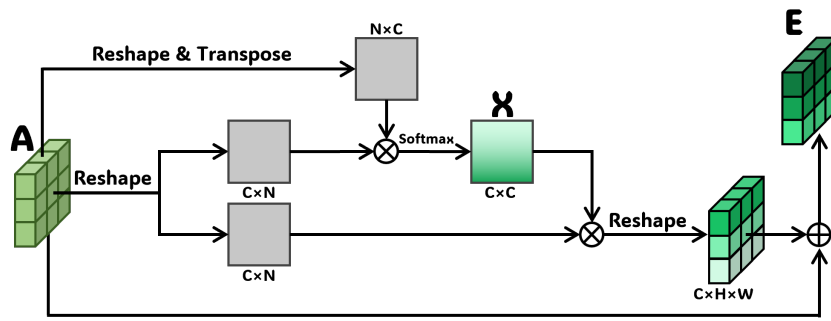


Figure 2. CAM structure diagram.

3.2. ACMIX

In the realm of person re-identification, researchers commonly employ two predominant methodologies for feature enhancement: CNN-attention and self-attention mechanisms. Although each approach offers distinct advantages, there exists a tendency among scholars to gravitate towards either one for the primary design of their models. However, it is noteworthy that self-attention, although capable of significantly expanding the receptive field of pedestrian images, demands a substantial corpus of training data. Conversely, CNN-attention, although characterized by its succinctness and practicality, often presents a limited receptive field compared to self-attention. In light of these considerations, we propose the incorporation of a Mix branch, which intricately integrates CNN-attention and self-attention modules through a shared procedural framework. This integration aims to capitalize on the strengths of both methods. The specific process is illustrated in Figure 3.

In the initial phase, the feature map undergoes projection through three 1×1 convolutional blocks. Subsequently, it is reshaped into N blocks, resulting in an enriched intermediate feature set of $3 \times N$.

In the second phase, in the self-attention branch, the intermediate feature set is consolidated into N groups, each containing three feature maps. Each feature within the group is obtained through a 1×1 convolution. The three respective feature maps are employed as queries, keys, and values, with the calculation formula as follows:

$$g_{ij} = \parallel_{l=1}^N \left(\sum_{a,b \in N_k(i,j)} A(q_{ij}^{(l)}, k_{ab}^{(l)}) v_{ab}^{(l)} \right) \quad (3)$$

where $\parallel$ denotes the concatenation of outputs from N attention heads, whereas $q_{ij}^{(l)}$, $k_{ab}^{(l)}$, and $v_{ab}^{(l)}$ denote the projection matrices for queries, keys, and values, respectively. $N_k(i, j)$ signifies the local pixel features within a spatial range of k centered at (i, j) . $A(q_{ij}^{(l)}, k_{ab}^{(l)})$ denotes the corresponding attention weight values for the internal region $N_k(i, j)$, with the weight assigned to values primarily determined by the degree of match between queries and keys. In the convolutional branch, considering a convolutional kernel $K \in \mathbb{R}^{C_{out} \times C_{in} \times k \times k}$, where K is the kernel size, and C_{in} and C_{out} represent the sizes of input and output channels. Let tensors $F \in \mathbb{R}^{C_{in} \times H \times W}$ and $G \in \mathbb{R}^{C_{out} \times H \times W}$ denote the input and output feature maps, where H and W represent height and width. Considering $f_{ij} \in \mathbb{R}^{C_{in}}$ and $g_{ij} \in \mathbb{R}^{C_{out}}$ as feature tensors for the corresponding pixels in F and G , respectively, the first phase can be represented as follows:

$$\tilde{g}_{ij}^{(p,q)} = K_{p,q} f_{ij} \quad (4)$$

where $K_{(p,q)} \in \mathbb{R}^{C_{out} \times C_{in}}$, $p, q \in \{0, 1, \dots, k-1\}$, represents the kernel weight values relative to the convolutional kernel position (p, q) . To simplify the formula, we define the operation $\tilde{f} \triangleq \text{Shift}(f, \Delta x, \Delta y)$ as follows:

$$\tilde{f}_{i,j} = f_{i+\Delta x, j+\Delta y}, \forall i, j, \quad (5)$$

where Δx and Δy denote the horizontal and vertical displacements, respectively. We employ a fully connected layer with dimensions $3N \times k^2N$ to generate k^2 feature maps in total, distributed across N groups. The generated feature maps undergo translation, and the translation formula is expressed as follows:

$$g_{ij}^{(p,q)} = \text{Shift}(\tilde{g}_{ij}^{(p,q)}, p - \lfloor k/2 \rfloor, q - \lfloor k/2 \rfloor) \quad (6)$$

The aggregation formula is expressed as follows:

$$g_{ij} = \sum_{p,q} g_{ij}^{(p,q)} \quad (7)$$

Hence, as evident from the above formulas, we initially convolve the input features to gather comprehensive information about individual images from local receptive fields. Finally, the outputs from both branches are combined, and their weighting is regulated by two scalar values, φ and ω . The formula is expressed as follows:

$$F_{\text{out}} = \varphi F_{\text{att}} + \omega F_{\text{conv}} \quad (8)$$

where F_{att} represents the result from the self-attention branch, and F_{conv} denotes the outcome from the convolutional branch.

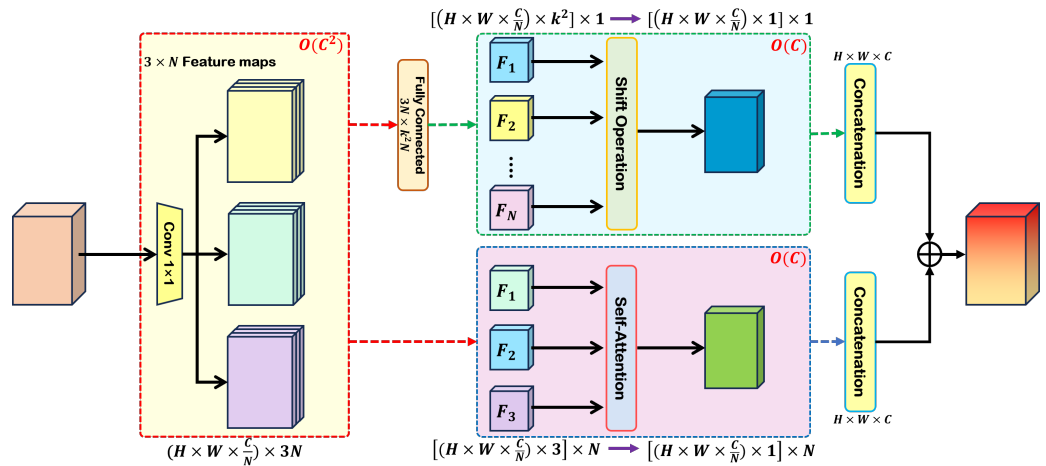


Figure 3. ACMIX structure diagram.

3.3. DPM

To obtain more discriminative features, we designed a dedicated Discriminative Part Mask (DPM) branch aimed at extracting discriminative information from non-core regions. As illustrated in Figure 4, through experimentation and observation, we discovered that areas outside the core regions of the human body (head, torso, or legs) still contain discriminative features [38], such as handheld items or body accessories. Consequently, we leverage the features acquired through the mixed attention branch to guide the suppression branch. This procedural step involves eliminating the most distinguishing areas of the person's image, thereby compelling the network to derive informative features from the remaining regions.

Specifically, we initially determine the maximum coordinates of the feature F_{mix} from the mixed attention branch, as follows:

$$x_c, y_c = \arg \max_{x,y} F_{\text{mix}} \quad (9)$$

where x_c and y_c represent the central coordinates of the region to be erased. To precisely determine the final size of the region to be erased, we conducted detailed comparisons in subsequent experiments. Ultimately, we determined the erased region F_e as follows:

$$F_e = \begin{cases} 0, & i \in (x_c \pm \frac{w}{12}), y \in (y_c \pm \frac{h}{4}) \\ F_{ij}, & \text{otherwise} \end{cases} \quad (10)$$

where w and h represent the width and height of the original image, respectively.

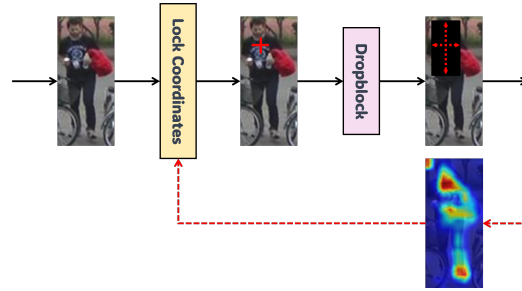


Figure 4. DPM structure diagram. The red dashed line represents the features F_{mix} output from the Mix branch, which are utilized for guidance.

3.4. LOSS

To better train MIX-Net, we employ a loss function comprising cross-entropy loss [8], triplet loss [1], and penalty terms for feature (O.F.) and weight (O.W.) regularization [39]. The formula is expressed as follows:

$$L = L_{xe} + \beta_{tri} L_{triplet} + \beta_{O.F.} L_{O.F.} + \beta_{O.W.} L_{O.W.} \quad (11)$$

where L_{xe} represents the cross-entropy loss, $L_{triplet}$ represents the triplet loss, $L_{O.F.}$ and $L_{O.W.}$ represent the respective penalty terms for feature and weight regularization, and $\beta_{[\cdot]}$ denotes the hyperparameters controlling the weights of the different loss functions.

4. Experiments and Discussion

4.1. Implementation Details

During the training process, all person images were resized to 384×128 and subjected to image augmentation techniques, including normalization, horizontal flipping, and random erasing. We utilized a ResNet-50 backbone network pre-trained on ImageNet and fine-tuned it using two transfer-learning algorithms. Hyperparameters were set as follows: $\beta_{tri} = 10^{-1}$, $\beta_{O.F.} = 10^{-6}$, $\beta_{O.W.} = 10^{-3}$, and α for triplet loss was set to 1.2. We employed an RTX 4090 GPU with 24 GB VRAM as the computing accelerator, set the number of epochs to 60, and batch size to 64, where each batch contained 16 identity IDs, with each ID comprising four instances. The Adam optimizer was employed, starting with a learning rate of 3×10^{-4} , which was reduced to 3×10^{-5} after 30 epochs and further decreased to 3×10^{-6} at 40 epochs.

Additionally, in accordance with prior research, we employed rank-1 (R1) accuracy and mean Average Precision (mAP) as our evaluation metrics.

4.2. Comparison Results

Table 1 showcases the comparative performance of our approach across three prominent person re-identification datasets: Market-1501, DukeMTMC, and MSMT17, in comparison to the latest state-of-the-art algorithms. From Table 1, it is observed that our algorithm outperforms most mainstream methods, achieving the best results in both mAP and rank-1 metrics on Market-1501, surpassing the second-best method by 1.0% and 0.2%, respectively. Similarly, on DukeMTMC, our algorithm achieves the best mAP and rank-1 metrics, surpassing the second-best method by 0.4% and 0.1%. On the most challenging

MSMT17 dataset, our algorithm achieves the best mAP and rank-1 metrics, significantly outperforming the second-best method by 5.0% and 2.8%. These results demonstrate that MIX-Net exhibits a highly competitive performance.

Table 1. Experimental results on Market1501, DukeMTMC, and MSMT17 datasets for various methods. Bold text represents the best results, whereas underlined text indicates the second-best results.

Methods	Market1501		DukeMTMC		MSMT17	
	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1
PCB+RPP [40]	81.6	93.8	69.2	83.3	40.4	68.2
PDC [20]	63.4	84.1	-	-	29.7	58.0
GLAD [41]	73.9	89.9	-	-	34.0	61.4
EDAGAN [42]	64.5	85.4	51.9	74.2	-	-
CDNet [43]	86.0	95.1	73.9	86.7	48.5	73.7
OSNet [44]	84.9	94.8	73.5	88.6	52.9	78.7
GASM [45]	84.7	95.3	74.4	88.3	52.5	79.5
IANet [46]	83.1	94.4	73.4	87.1	46.8	<u>75.5</u>
BAT-net [47]	84.7	95.1	77.3	87.7	56.8	79.5
JDGL [48]	86.0	94.8	74.8	86.6	<u>52.3</u>	<u>77.2</u>
PAT [49]	88.0	95.4	78.2	88.8	-	-
DRL-Net [50]	<u>86.9</u>	94.7	76.6	<u>88.1</u>	55.3	78.4
MGN [51]	86.9	<u>95.7</u>	<u>78.4</u>	88.7	52.1	76.9
IGOAS [52]	84.1	<u>93.4</u>	<u>75.1</u>	86.8	-	-
MIX-Net	89.0	95.9	78.8	88.9	61.8	82.3

Table 2 showcases the experimental findings pertaining to occluded scenarios, showcasing the consistent and formidable performance of MIX-Net even amidst intricate scenarios characterized by occlusions. This remarkable resilience can be primarily attributed to the robustness instilled by our meticulously crafted hybrid attention mechanism. Furthermore, the Discriminative Part Mask (DPM) adeptly supplements missing information, thereby bolstering MIX-Net's robustness and fortifying its resistance against interference. Consequently, MIX-Net exhibits remarkable proficiency in navigating through diverse and challenging scenarios with aplomb.

Table 2. Experimental results on Occluded-Duke and Occluded-REID datasets for various methods. Bold text represents the best results, whereas underlined text indicates the second-best results.

Method	Occluded-Duke		Occluded-REID	
	mAP	Rank-1	mAP	Rank-1
DSR [53]	30.4	40.8	62.8	72.8
SFR [54]	32.0	42.3	-	-
FPR [55]	-	-	68.0	78.3
HOReID [17]	43.8	55.1	70.2	80.3
CBDB-Net [56]	<u>38.9</u>	50.1	-	-
PAFM [57]	42.3	55.1	68	76.4
ASAN [58]	<u>43.8</u>	55.4	<u>71.8</u>	<u>86.8</u>
PGFA [35]	37.3	<u>51.4</u>	-	-
MIX-Net	52.4	60.7	79.1	86.9

In order to investigate the body part information focused on by MIX-Net and other networks, we employed Grad-cam [59] to generate heatmap visualizations in Figure 5. Analysis of the heatmap reveals that ResNet and IGOAS emphasize rich global information. However, this approach often weakens the impact of key regional information on the final discriminative features. Conversely, OSNet and PCB exhibit a penchant for focusing

attention on the most discriminative key regions of the body. Nonetheless, this inclination may engender a concentration of final features on a restricted body area, potentially leading to a dearth of diversity in the extracted features. In stark contrast, MIX-Net adeptly navigates a balance, adeptly capturing information from pivotal body parts while concurrently exploring latent cues from other less salient yet nonetheless critical regions. Consequently, the resultant features extracted by MIX-Net encapsulate information from key regions with substantial weight proportions, alongside potentially invaluable cues from less central regions with diminished weight proportions.

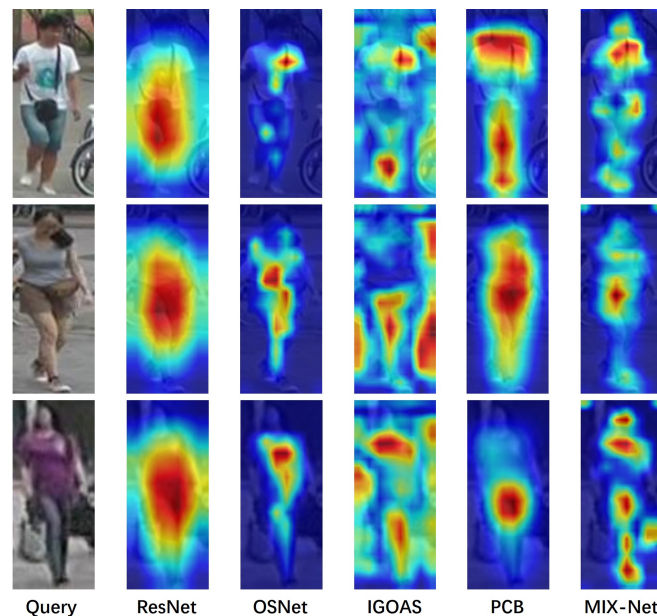


Figure 5. Heatmap visualization of comparative experiments. On the Market1501 dataset, we conducted heatmap visualization for comparative experiments. The red areas represent the focal regions attended to by the models, whereas the blue areas denote less significant regions.

For a more comprehensive and intuitive evaluation of MIX-Net’s performance, we selected three sets of pedestrian images, representing front-view, back-view, and side-view perspectives from left to right, as shown in Figure 6. In comparison to other state-of-the-art models, MIX-Net demonstrates consistently robust performance, particularly in distinguishing pedestrians with similar attributes.

In more granular detail, both PCB and ResNet demonstrate a notable reduction in accuracy when confronted with variations in perspective or blurred pedestrian images. This decline primarily arises from their ability to focus on pivotal body parts and extract features, coupled with their susceptibility to interference from extraneous information beyond the human body contour, thus resulting in diminished stability. Conversely, OSNet exhibits robust discriminative capabilities for pedestrians with comparable or identical viewpoints. Nonetheless, its pronounced emphasis on key body areas at the expense of global information extraction renders OSNet less adept at sustaining optimal performance in intricate scenarios characterized by perspective variations. IGOAS prioritizes the acquisition of diverse information and yields commendable outcomes under normal circumstances. However, its overemphasis on global information tends to diminish the significance of critical regions, thereby impeding stability when confronted with pedestrians sharing similar attributes such as clothing or body shapes. This propensity causes IGOAS to struggle to maintain stability when confronted with pedestrians sharing similar attributes, such as similar clothing or body shapes. MIX-Net, owing to its adept hybrid attention mechanism design and the incorporation of DPM for latent information extraction, excels in rank-5 experiments. It consistently upholds robustness and stability, whether dealing with standard pedestrian images or navigating through complex scenarios involving pedestrians.



Figure 6. Visualization of rank-5 comparison experiments with other mainstream models. On the Market1501 dataset, we conducted visualization of rank-5 rankings in comparative experiments, where red borders represent misidentifications and green borders represent correct identifications.

To further investigate the performance of MIX-Net and other networks in the presence of occluded body parts, we present heatmap visualizations in Figure 7, showcasing scenarios where crucial body regions are partially concealed. Analyzing the images reveals that MIX-Net maintains robust performance in the face of complex occlusion, primarily attributed to the potent anti-interference capabilities provided by the mixed attention mechanism in the Mix branch. Additionally, the DPM contributes by offering latent information to assist the network in recognizing targets under challenging conditions. In contrast, other networks struggle to balance the extraction of diverse information and crucial region details, resulting in a significant degradation in performance under complex scenarios.

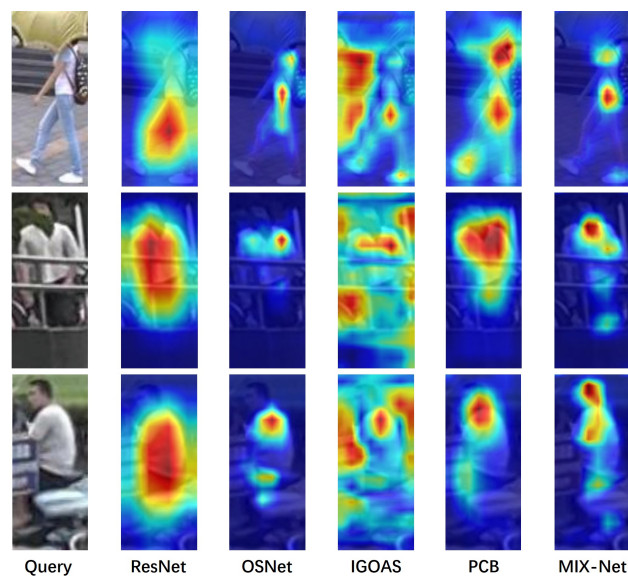


Figure 7. Heatmap visualization experiment with occluded targets. We conducted heatmap visualizations, focusing on scenarios with occluded targets. In the visualizations, the red areas denote regions where the model emphasizes important information, whereas the blue areas represent less critical regions.

In summary of the aforementioned experiments, ResNet, despite its strong generality for various computer vision tasks, lacks precision for individual person re-identification tasks. The features it extracts contain a considerable amount of redundant information, resulting in large high-attention regions in the heatmaps. This characteristic makes it prone to overlooking or misidentifying individuals with similar attributes in rank-5 experiments. On the contrary, OSNet, tailored explicitly for person re-identification, adeptly aligns with the human body contour. However, it demonstrates weaker capabilities in extracting information from occluded or less significant areas such as handheld items, limbs, and low-light regions. This inadequacy manifests in the heatmaps as a concentration of attention on localized areas, resulting in diminished discrimination for individuals with less attention in highlighted regions during rank-5 experiments. IGOAS adopts a design philosophy that leans toward extracting features from all global information available to a person. Although this network structure enriches the final feature information, it also introduces excessive information that negatively impacts the model's performance. In the heatmaps, the attention areas are dispersed across the entire image, and in rank-5 experiments, IGOAS performs poorly when faced with relatively uniform background targets. PCB's design philosophy involves segmenting and extracting features horizontally from distinct body parts before amalgamating them. This approach allows the network to exploit relationships between body parts and extract diverse information meticulously. However, the segmented and merged architecture of PCB fails to precisely identify the globally critical area, leading to instability in performance. Although PCB demonstrates proficiency in extracting good features in some instances, its accuracy significantly diminishes in the presence of interference or occlusion. Furthermore, in heatmap experiments, PCB displays inconsistent performance across several targets and struggles to accurately conform to the human body contour. During rank-5 experiments, PCB encounters challenges in accurately identifying targets with varying perspectives.

In contrast, MIX-Net benefits from the hybrid attention mechanism that combines self-attention and CNN-attention. The CNN-attention allows MIX-Net to focus on the core areas of the human body, whereas advanced self-attention enables the network to have a strong focus on aligning with the human body contour. Meanwhile, DPM serves as a complement to the attention mechanisms, extracting latent information. In heatmap experiments, MIX-Net can simultaneously pinpoint multiple important body parts to extract crucial information, while also focusing on other secondary important parts and regions to extract discriminative latent information. In rank-5 experiments, MIX-Net demonstrates robust performance and strong capabilities.

4.3. Ablation Studies

To comprehensively investigate the impact of different modules and parameters in our designed network on the overall performance, we present the results of ablation experiments conducted in Table 3.

Table 3. Ablation experiments of MIX-Net conducted on the Market1501 dataset, with bold highlighting the optimal results.

CAM	O.F.	O.W.	ACMIX	DPM	$L_{triplet}$	Rank-1	mAP
						90.1	76.1
✓						92.2	78.4
	✓					92.0	82.0
		✓				92.6	78.3
	✓	✓				92.3	81.9
✓	✓	✓				93.8	85.1
✓	✓	✓	✓			94.2	86.1
✓	✓	✓	✓		✓	95.1	88.0
✓	✓	✓	✓	✓		94.5	87.4
✓	✓	✓	✓	✓	✓	95.9	89.0

From Table 3, it can be observed that using CAM and the hybrid attention module ACMIX individually can improve the network's performance. The combination of these three attention mechanisms further enhances the network's capabilities, demonstrating the complementary effects of the two attention mechanisms employed. The use of O.F. and O.W. also improves the network's performance, and their combination achieves better results, validating the effectiveness of the approach. The DPM, when integrated, significantly boosts mAP and shows improvement in rank-1. Additionally, combining DPM with other attention mechanisms yields superior results. Lastly, the addition of the triplet loss contributes to the best overall performance.

In the experimental design, we observed that the size setting of the erased region in the DPM significantly influences its capability to extract latent information, consequently affecting the overall network performance. Therefore, in Figure 8, we illustrate the impact of the erased region size in the inhibition module on the mAP and rank-1 metric of MIX-Net. Experimental results indicate that setting the erased region of the inhibition module to a height of 3/6 and a width of 1/6 achieves the highest accuracy. Experimental results indicate that DPM does indeed have an impact on the overall network. In comparison to the relatively stable rank-1 metric, the mAP is more significantly affected. This discrepancy arises because DPM is designed to excavate latent information in secondary important regions to assist the network in recognizing challenging target samples. This enhancement in recognition capability for complex targets leads to an improvement in average precision.

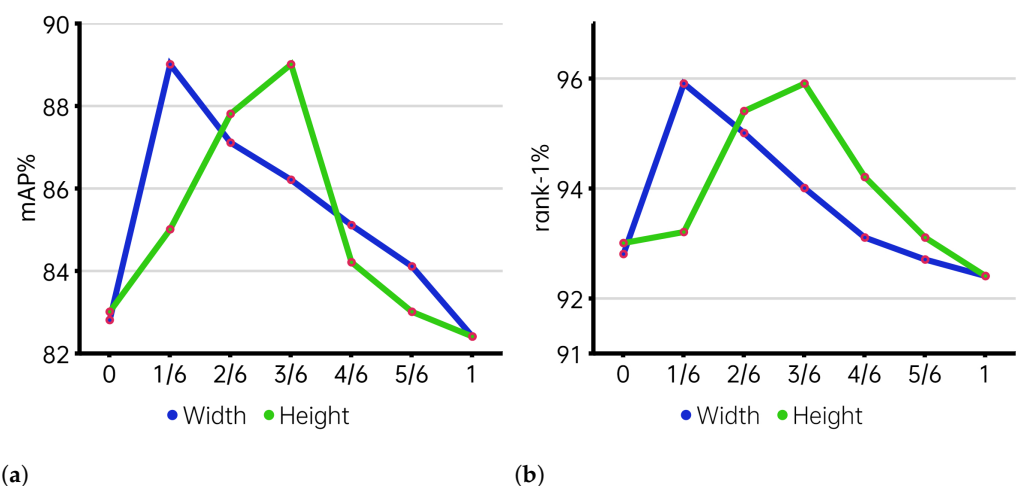


Figure 8. DPM parameter ablation experiments were conducted on the Market1501 dataset. (a) The impact of the erased region size on mAP. (b) The impact of the erased region size on rank-1.

However, the influence of DPM on rank-1 performance is limited, as rank-1 heavily depends on the information excavation capability of key regions for recognition. Therefore, although DPM provides assistance, its impact on rank-1 is present but constrained.

To further investigate the discriminative capabilities of our different modules for pedestrians with similar attributes, we conducted visualization experiments, as shown in Figure 9. CAM and ACMIX significantly enhance the network's discriminative ability for pedestrian attributes. Additionally, DPM demonstrates a notable improvement in discerning pedestrians with similar attributes, such as similar clothing or body shapes.



Figure 9. Rank-5 visualization of ablation experiments. Ablation experiments with rank-5 visualization were conducted on the Market1501 dataset. In the visualizations, red borders represent misidentified identities, whereas green borders indicate correct identities.

5. Conclusions

The present study introduces an innovative MIX-Net architecture designed to learn diverse person features by simultaneously focusing on both salient and latent regions. Leveraging an improved backbone network, ResNet-50, for fundamental feature extraction, the features undergo enhancement through attention, suppression, and global branches. The training process involves the utilization of cross-entropy loss and triplet loss. Extensive comparative experiments validate the superior performance of MIX-Net, with ablative studies demonstrating substantial contributions from each constituent module to the overall performance. In our future endeavors, we aim to enhance the DPM by incorporating an attention module tailored to augment the discriminative information in secondary regions. Additionally, we seek to refine the fusion approach between DPM and the attention branch through dynamic aggregation, thereby further enhancing the performance of DPM and enabling the model to adapt to more complex and constrained environments. Moreover, we aspire to extend the concept of hybrid attention to a broader spectrum of person re-identification tasks involving intricate scenarios.

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

Funding: This work is supported by: (1) Research on Key Technologies for 3D Point Cloud Object Recognition in Industrial Sorting Robots: supported by the Applied Basic Research Project of the Science and Technology Department of Liaoning Province, (no. 2022JH2/101300274); (2) Exploration of an Integrated Practice and Innovation Capability Cultivation Model for Professional Master's Students with a Six-in-One and Industry-Education Integration Approach: funded by the Liaoning Provincial Graduate Education Teaching Reform Research Project for the Year 2023 (no. LNYJG2023117); (3) Study on a Vital Signs Perception Model Based on the Fresnel Zone in a Non-contact Environment: supported by the General Research Project (Surface Project) of the Education Department of Liaoning Province in 2022, (no. LJKMZ20220676).

Institutional Review Board Statement: Not applicable.

Data Availability Statement: The data presented in this study are available upon request from the corresponding author. The data are not publicly available due to only being available to teams interested in collaboration.

Conflicts of Interest: The authors declare that they have no known competing financial interest or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Hermans, A.; Beyer, L.; Leibe, B. In defense of the triplet loss for person re-identification. *arXiv* **2017**, arXiv:1703.07737.
2. 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](#)]
3. Li, Z.; Chang, S.; Liang, F.; Huang, T.S.; Cao, L.; Smith, J.R. Learning locally-adaptive decision functions for person verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3610–3617.
4. 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.
5. Dutta, S.; Basarab, A.; Georgeot, B.; Kouamé, D. DIVA: Deep unfolded network from quantum interactive patches for image restoration. *arXiv* **2022**, arXiv:2301.00247.
6. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the integration of self-attention and convolution. In Proceedings of the IEEE/CVF conference On Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 815–825.
7. Wu, L.; Shen, C.; Hengel, A.v.d. Personnet: Person re-identification with deep convolutional neural networks. *arXiv* **2016**, arXiv:1601.07255.
8. Zhou, K.; Xiang, T. Torchreid: A Library for Deep Learning Person Re-Identification in Pytorch. *arXiv* **2019**, arXiv:1910.10093.
9. 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; pp. 5098–5107.
10. 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.
11. 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, IEEE, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
12. Yi, D.; Lei, Z.; Liao, S.; Li, S.Z. Deep metric learning for person re-identification. In Proceedings of the 2014 22nd International Conference on Pattern Recognition, IEEE, Stockholm, Sweden, 24–28 August 2014; pp. 34–39.
13. Ahmed, E.; Jones, M.; Marks, T.K. An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3908–3916.
14. 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.
15. Wang, Y.; Chen, Z.; Wu, F.; Wang, G. Person re-identification with cascaded pairwise convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1470–1478.
16. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3219–3228.
17. Wang, G.; Yang, S.; Liu, H.; Wang, Z.; Yang, Y.; Wang, S.; Yu, G.; Zhou, E.; Sun, J. High-order information matters: Learning relation and topology for occluded person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6449–6458.
18. Deng, W.; Zheng, L.; Ye, Q.; Kang, G.; Yang, Y.; Jiao, J. Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 994–1003.
19. Liu, C.; Chang, X.; Shen, Y.D. Unity style transfer for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 6887–6896.
20. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
21. Liu, J.; Ni, B.; Yan, Y.; Zhou, P.; Cheng, S.; Hu, J. Pose transferrable 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. 4099–4108.
22. Mirza, M.; Osindero, S. Conditional generative adversarial nets. *arXiv* **2014**, arXiv:1411.1784.
23. Tian, X.; Zhang, Z.; Lin, S.; Qu, Y.; Xie, Y.; Ma, L. Farewell to mutual information: Variational distillation for cross-modal person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 1522–1531.
24. Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; Zhang, Z. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 842–850.
25. 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.

26. Li, S.; Bak, S.; Carr, P.; Wang, X. Diversity regularized spatiotemporal attention for video-based 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. 369–378.
27. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
28. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 365–381.
29. 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; pp. 8351–8361.
30. Zhuo, J.; Chen, Z.; Lai, J.; Wang, G. Occluded person re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), IEEE, San Diego, CA, USA, 23–27 July 2018; pp. 1–6.
31. Liu, J.; Zha, Z.J.; Hong, R.; Wang, M.; Zhang, Y. Deep adversarial graph attention convolution network for text-based person search. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 665–673.
32. 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.
33. 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; Springer: Berlin/Heidelberg, Germany, 2016; pp. 17–35.
34. Wei, L.; Zhang, S.; Gao, W.; Tian, Q. Person transfer gan to bridge domain gap for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 79–88.
35. 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; pp. 542–551.
36. 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.
37. Fu, J.; Liu, J.; Tian, H.; Li, Y.; Bao, Y.; Fang, Z.; Lu, H. Dual attention network for scene segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3146–3154.
38. Yu, Y.; Yang, S.; Hu, H.; Chen, D. Attention-Guided Multi-Clue Mining Network for Person Re-identification. *Neural Process. Lett.* **2022**, *54*, 3201–3214. [[CrossRef](#)]
39. Kalayeh, M.M.; Basaran, E.; Gökmen, M.; Kamasak, M.E.; Shah, M. Human semantic parsing 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. 1062–1071.
40. 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.
41. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: Global-local-alignment descriptor for pedestrian retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 420–428.
42. Wang, Y.; Jiang, K.; Lu, H.; Xu, Z.; Li, G.; Chen, C.; Geng, X. Encoder-decoder assisted image generation for person re-identification. *Multimed. Tools Appl.* **2022**, *81*, 10373–10390. [[CrossRef](#)]
43. 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, Nashville, TN, USA, 20–25 June 2021; pp. 6729–6738.
44. 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.
45. He, L.; Liu, W. Guided saliency feature learning for person re-identification in crowded scenes. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020, Proceedings, Part XXVIII* 16. Springer: Berlin/Heidelberg, Germany, 2020; pp. 357–373.
46. 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.
47. Fang, P.; Zhou, J.; Roy, S.K.; Petersson, L.; Harandi, M. Bilinear attention networks for person retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8030–8039.
48. Zheng, Z.; Yang, X.; Yu, Z.; Zheng, L.; Yang, Y.; Kautz, J. Joint discriminative and generative learning for person re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2138–2147.
49. 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, Nashville, TN, USA, 20–25 June 2021; pp. 2898–2907.

50. Jia, M.; Cheng, X.; Lu, S.; Zhang, J. Learning disentangled representation implicitly via transformer for occluded person re-identification. *IEEE Trans. Multimed.* **2022**, *25*, 1294–1305. [\[CrossRef\]](#)
51. 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; pp. 274–282.
52. Zhao, C.; Lv, X.; Dou, S.; Zhang, S.; Wu, J.; Wang, L. Incremental generative occlusion adversarial suppression network for person ReID. *IEEE Trans. Image Process.* **2021**, *30*, 4212–4224. [\[CrossRef\]](#) [\[PubMed\]](#)
53. He, L.; Liang, J.; Li, H.; Sun, Z. Deep spatial feature reconstruction for partial person re-identification: Alignment-free approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7073–7082.
54. He, L.; Sun, Z.; Zhu, Y.; Wang, Y. Recognizing partial biometric patterns. *arXiv* **2018**, arXiv:1810.07399.
55. He, L.; Wang, Y.; Liu, W.; Zhao, H.; Sun, Z.; Feng, J. Foreground-aware pyramid reconstruction for alignment-free occluded person re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8450–8459.
56. Tan, H.; Liu, X.; Bian, Y.; Wang, H.; Yin, B. Incomplete descriptor mining with elastic loss for person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 160–171.
57. Yang, J.; Zhang, C.; Tang, Y.; Li, Z. PAFM: Pose-drive attention fusion mechanism for occluded person re-identification. *Neural Comput. Appl.* **2022**, *34*, 8241–8252. [\[CrossRef\]](#)
58. Jin, H.; Lai, S.; Qian, X. Occlusion-sensitive person re-identification via attribute-based shift attention. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2170–2185. [\[CrossRef\]](#)
59. 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; pp. 618–626.

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.