electronics

MDPI

Article

An Orientation-Aware Attention Network for Person Re-Identification

Dongshu Xu [1,2], Jun Chen [1,2,*] and Xiaoyu Chai [1,2]

[1] National Engineering Research Center for Multimedia Software, School of Computer, Wuhan University, Wuhan 430072, China; xudongshu@whu.edu.cn (D.X.); stevenchai@whu.edu.cn (X.C.)
[2] Hubei Key Laboratory of Multimedia and Network Communication Engineering, Wuhan University, Wuhan 430072, China
* Correspondence: chenj.whu@gmail.com

Abstract: Humans always identify persons through their characteristics, salient attributes, and these attributes' locations on the body. Most person re-identification methods focus on global and local features corresponding to the former two discriminations, cropping person images into horizontal strips to obtain coarse locations of body parts. However, discriminative clues corresponding to location differences cannot be discovered, so persons with similar appearances are often confused because of their alike components. To address the above problem, we introduce pixel-wise relative positions for the invariance of their orientations in viewpoint changes. To cope with the scale change of relative position, we combine relative positions with self-attention modules that perform on multi-level features. Moreover, in the data augmentation stage, mirrored images are given new labels due to the conversion of the relative position along a horizontal orientation and change in visual chirality. Extensive experiments on four challenging benchmarks demonstrate that the proposed approach shows its superiority and effectiveness in discovering discriminating features.

Keywords: person re-identification; orientation-aware attention; visual chirality

check for updates

Citation: Xu, D.; Chen, J.; Chai, X. An Orientation-Aware Attention Network for Person Re-Identification. *Electronics* **2024**, *13*, 910. https://doi.org/10.3390/electronics13050910

Academic Editor: George A. Papakostas

Received: 22 January 2024
Revised: 11 February 2024
Accepted: 26 February 2024
Published: 27 February 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

In recent years, with the development and progress of smart cities, public security has become increasingly significant. Personal trajectories play an important role in security. In smart cities, the urban monitoring system has a large amount of cameras that generate massive video data, which can be used for person identification and tracking. Face recognition and digital identity have been widely used for person identity acquisition. However, due to the low resolution and different views of person images captured in surveillance cameras, face recognition can be ineffective in most scenes. It is necessary to integrate overall characteristics to capture enough clues for person–image matching. Thus, person re-identification (ReID) has become an important task in video surveillance.

Person ReID is an image retrieval task, aiming to associate pedestrian images captured by non-overlapping cameras [1]. In the general procedure of person ReID, given a query image from one camera and gallery images from other cameras, each image is transformed into a feature embedding, and then feature similarities are ranked between those embeddings of query and gallery images [2]. Those images with the same identities should rank forefront. Hence, identifying robust discriminative features is a crucial factor in person ReID.

Tremendous improvements [3–9] have been achieved in recent years by studying many practical problems to obtain discriminative features, e.g., viewpoint change, pose variation, and occlusion. Commonly, they use a main backbone to obtain global features and crop person images into horizontal strips to extract local features. However, cropping images using a fixed interval brings misalignments of local features, since some person images are acquired with inaccurate detection boxes, such as boxes with the person not

centered or boxes with partial bodies. Therefore, the attention scheme [10–12] has been introduced to enforce the model to capture cardinal discriminative local features, which boosts the performance of person ReID models greatly. These methods usually focus on the existence of discriminative patterns without regard for positions and orientations. However, persons with similar appearances usually have similar patterns on their clothing, such as logos, figures, etc. These kinds of similar appearances can hardly be discriminated by these methods, although humans can usually distinguish them.

Humans always identify a person through their characteristics, salient attributes, and these attributes' locations on the body. The aforementioned person ReID methods almost always focus on the former two discriminations and have remarkable performance. Recently, pose estimation methods [13] have been introduced to localize body parts to enhance the correspondences between local features and their positions. However, there exists a domain gap between the pose estimation dataset and the person ReID dataset. It is hard to accurately localize the positions of body parts, which restricts the development of pose-guided person ReID methods. Most studies [14,15] crop images into horizontal strips to obtain coarse vertical position information, which need plenty of memory and are vulnerable to being influenced by misalignment. Although these part localization methods provide clues for person ReID, the generated inaccurate and coarse positions cannot help discover fine-grained position-aware discriminative clues.

To obtain stable and effective position information, we analyze the position variation when changing viewpoints and fixing other variables. We discover that the relative position between visible parts has some characteristics. In circular views, since only half of the visible part of the whole body can be captured, the orientations for horizontal and vertical components of relative positions between visible patterns are invariant. As shown in Figure 1, the horizontal components of relative positions between the bag shown in yellow boxes and the white parallel lines shown in red boxes are invariant if the comparable patterns are visible. The vertical components are also invariant for persons in images captured in video surveillance, which are almost never upside-down. For the scale of relative position, it is mainly related to the distances from viewpoint to person location and the spin angle of body rotation. In practical scenarios, the distances between the camera and the pedestrians are much greater than the widths of pedestrians. We can suppose it is a nearly linear correlation between the relative position and the distance of the viewpoint and captured part. Hence, the impact of neighbor body parts can be defined as a linear form. Furthermore, visual chirality [16], an orientation phenomenon of images, also proves usefulness with regard to the aforementioned orientation clue for discrimination. Note that horizontal flipping with the same label, which is widely used in data augmentation, would probably bring disorder to the calculation of relative positions.

In this paper, we propose a novel Multi-level Position-aware Global Attention Network (MPGA-Net), which uses a simple but effective framework to introduce discriminative position information including global and local features with relative position encoding. Furthermore, given new labels to mirrored images, MPGA-Net treats the visual chirality of person images as a new clue to identify persons. To obtain the position-aware global and local relationships in different semantic levels, MPGA-Net inserts the position-aware attention module (PAM) into several residual blocks. Meanwhile, to balance the discriminating effects of feature and position, we proposed an Adaptive Label Smoothing strategy to ensure that features of mirrored images are closer to those of original images than images with other identities.

The sections of this article are arranged as follows:

Section 1 introduces the research background and significance of person ReID, then describes its definition and mainstream ideas, and finally elaborates on motivation and innovation points.

Section 2 introduces the development and innovation of person ReID algorithms based on position-aware representation and self-attention and then analyzes the shortcomings of existing algorithms. Finally, it describes the innovative improvement briefly.

Section 3 provides the framework of the orientation-aware person ReID algorithm, which includes re-labeling augmented images in preprocessing, multi-scale self-attention modules with position encoding in network construction, and an Adaptive Label Smoothing strategy. The idea for label regulation mainly corresponds to the orientation problems, and the module design and network construction are mainly related to the position-aware representation acquirement.

Section 4 introduces person ReID datasets, evaluation protocol, and experimental configuration. Then, we design comparative experiments and ablation studies and analyze the results. Meanwhile, the visualizations of the response map and ranking results are displayed to verify the effectiveness of the algorithm design.

Section 5 summarizes the content and the innovations of the paper and provides future research directions on analyses of the shortcomings of the proposed algorithm.

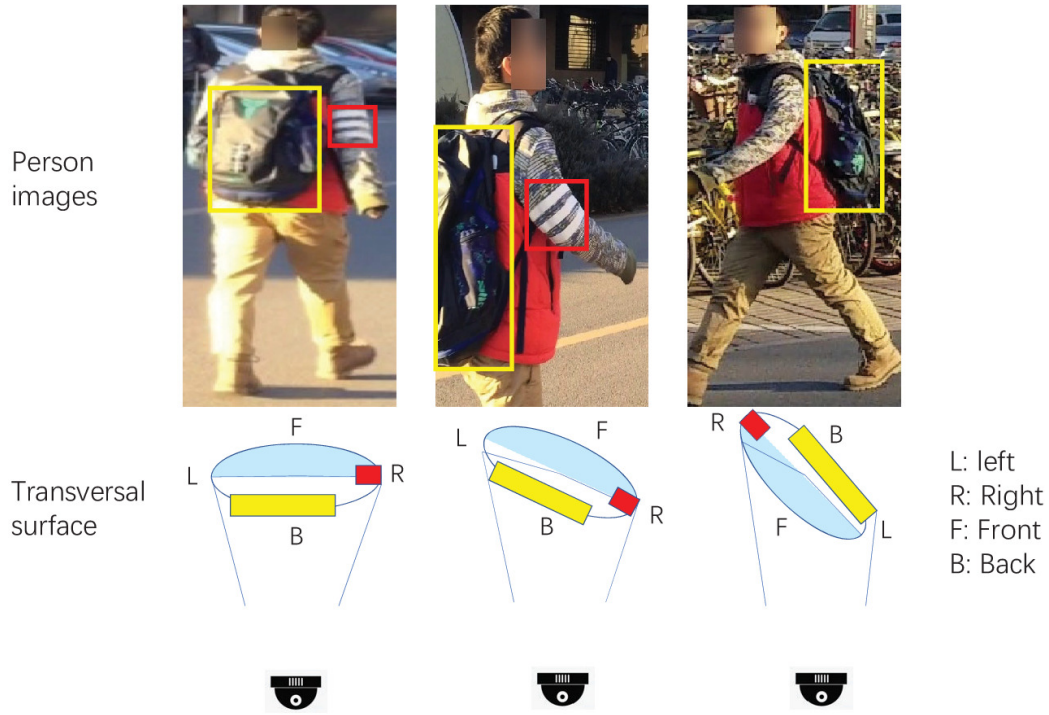

Figure 1. Example of relative position orientation between person areas. The orientation for the horizontal component of the relative position remains unchanged if person areas are visible.

2. Related Work

2.1. Re-Identification Based on Position-Aware Representation

Position information is usually applied in person ReID to obtain local part representations. Zhao et al. [17] proposed a part-aligned deep neural network that adopted a part map detector to localize aligned parts. The dynamic part alignment strategy provided feature-based position-aware representation. Sun et al. [14] proposed a Part-based Convolutional Baseline (PCB) that uniformly divided images into horizontal strips using a fixed interval, extracted part features from each strip, and then refined the consistency within parts. Position information by fixed partitioning strategy is coarse, and part features usually need to be refined or aligned.

To obtain more accurate parts, pose estimation is introduced into person ReID. Zhao et al. [18] proposed a part-based framework to integrate multi-level features of the human body structure. It formulates the body part areas through key points obtained by a pose estimation network. Suh et al. [19] adopted a two-stream network to obtain local part features, which combined appearance features in one stream and a part map in another stream by bilinear pooling. The pose estimation components are always trained on pose

estimation datasets and then directly used to generate body key points. There exist domain gaps between pose estimation datasets and person ReID datasets. The generated position information is unconfirmed.

Recently, transformers have been widely used in nearly every computer vision scene. The transformer also has the position-encoding stage due to the lack of position information in self-attention. He et al. [20] proposed a pure transformer-based person ReID framework, which integrates side information embeddings and a jigsaw patches module with a transformer to obtain robust features. Transformer-based methods always have higher performance on large datasets, but they usually need higher computational cost and many samples to train.

2.2. Re-Identification Based on Self-Attention

Attention [21] has been widely studied in recent years. Zhang et al. [22] proposed a relation-aware global attention module to strengthen feature discrimination in a global perception, which integrates local features and global relations to calculate the attention weight. Chen et al. [23] proposed a salience-guided cascaded suppression network to mine salient features and integrate them into the final representation in a cascaded manner. Self-attention-based methods usually adopt a module similar to a non-local block [24] to obtain attention to re-weight features. The attention calculation is based on affinity within local features. This manner may reduce the importance of pixel neighbors, which would weaken local information.

In this paper, we propose a novel module to learn a global representation with the accumulation of attentioned local features and their relative position clues to alleviate the problem of lacking position information in the self-attention module.

3. Proposed Method

3.1. Overall Architecture

Image-based person ReID aims at matching cross-camera person images pairwise.

Given a set of person images for training $X = \{x_i\}_{i=1}^{N}$ containing N samples from P pedestrians with their corresponding identity labels as $Y = \{y_i\}_{i=1}^{N}$, the goal is to explore discriminative features to identify persons.

To make features more discriminative, we propose a novel MPGA-Net to learn a global representation with the accumulation of attentioned local features and their relative position clues, as shown in Figure 2. In the data preparation stage, all images for training are flipped horizontally and given new identities, named the Horizontal Flipping with New Identities (HFNI) strategy, which is beneficial for discovering orientation clues for bilateral asymmetric parts. After horizontal flipping, if the source image x_i is labeled y_i, the mirrored image can be marked as x_{i+N} with the label $y_i + P$.

To enhance network extensibility, we design a plug-and-play module, named the Position-aware Attention Module (PAM), which can be easily inserted into any block of ResNet [25]. In this paper, the even sequence set of blocks in each layer has added the PAMs that exploit relative position embeddings and self-attention on multi-level feature maps. PAMs are placed after the second ReLU in the even blocks. In our view, the PAMs discover discriminative areas by global attention based on features and relative positions. After all residual layers, a Global Average Pooling (GAP) layer summarizes multi-level position-aware features to global features.

Due to the newly added mirrored images and identities, we propose an Adaptive Label Smoothing (ALS) strategy to force the network not only to discover position-aware clues but also to learn more discriminative representations. Finally, we adopt the cross-entropy loss with augmented new labels and the triplet loss with source old labels because most personal appearances are bilaterally symmetric.

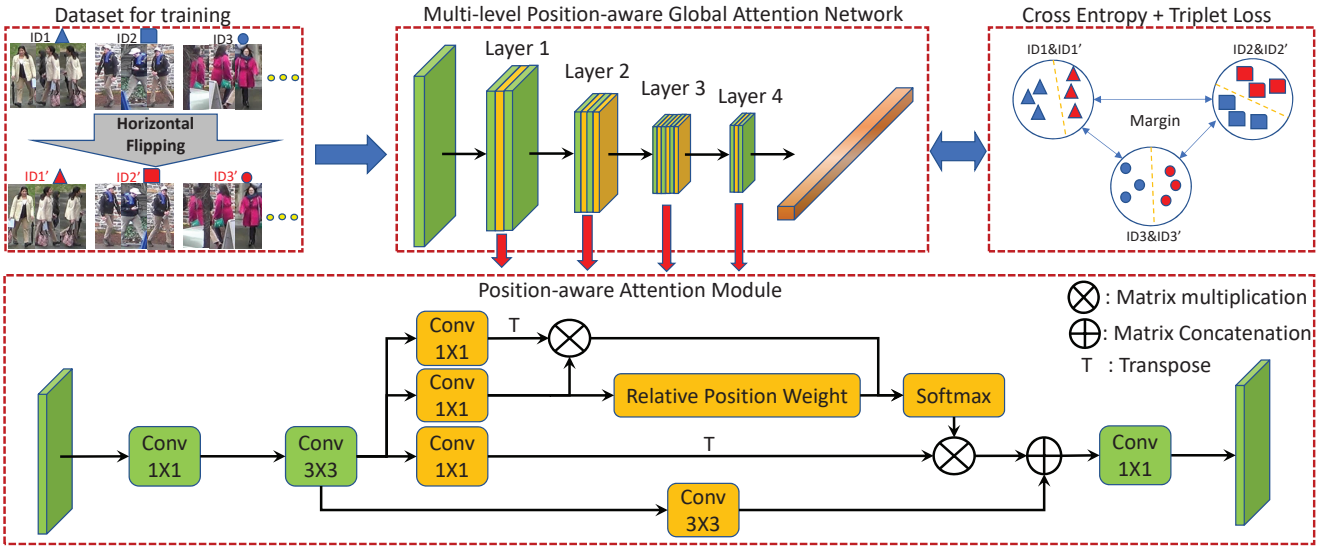

Figure 2. The overall architecture of MPGA-Net. The training dataset is built by source training images, and mirrored images are given new labels. The MPGA-Net is constructed by inserting PAMs into the even sequence set of residual blocks in each layer. The loss function is built on the cross-entropy loss and the triplet loss with an Adaptive Label Smoothing strategy.

3.2. Position-Aware Attention Module

The Position-aware Attention Module is developed for exploring position-aware spatial attention on multi-level feature maps. On one hand, most existing attention-based methods [22,26] in person ReID always place the attention module on the link of two layers, which blocks the residual propagation. On the other hand, existing approaches usually neglect the importance of the relative position of attentioned features. Motivated by the above observation, we design a position-aware module that can be inserted into each block to learn multi-level positional attention. Inspired by CBAM [27], as relative positions are the spatial correlation between areas, the PAM is built on a spatial attention module, as illustrated in Figure 2. Note that PAMs inserted in a bottleneck before the last 1×1 convolution can reduce computing complexity, since the channel number is quadrupled by the last convolution in each block. Moreover, 1×1 convolution combines features and relative positions.

Let $f_i \in \mathbb{R}^{h \times w \times c}$ denote the input feature map of the i-th image for PAM, where h, w and c are the height, width, and channels of the feature map, respectively. We use three 1×1 convolutions followed by Instance Normalization layers to obtain query embedding $Q \in \mathbb{R}^{h \times w \times 2c}$, key embedding $K \in \mathbb{R}^{h \times w \times 2c}$ and value embedding $V \in \mathbb{R}^{h \times w \times \frac{c}{2}}$, respectively. To obtain more clues of the relative position information, the channel number of query embedding is doubled as $2c$. Based on the non-local settings, the channel number of key embedding is the same as the query. To reduce computational complexity and preserve a block of channel space for local feature aggregation, the channel number of value embedding is reduced to $c/2$.

According to [28], relative position can be resolved into an accumulation of relative height and relative width. The accumulation of relative height and width can be calculated by query embedding Q and learnable Relative Position Weight (RPW). The Position-aware Global Attention (PGA) is computed as follows:

$$PGA = Softmax(\frac{QK^T + S_H + S_W}{\sqrt{2c}})V \quad (1)$$

where S_H and S_W are relative height logits and relative width logits, respectively.

$$S_H[t,k] = q_t^T r^H_{k_y - t_y}, S_W[t,k] = q_t^T r^W_{k_x - t_x} \tag{2}$$

where t and k are pixels, q_t is the query vector of pixel t, and $r^H_{k_y - t_y}$ and $r^W_{k_x - t_x}$ are learned embeddings for relative height $k_y - t_y$ and relative width $k_x - t_x$, as shown in Figure 3.

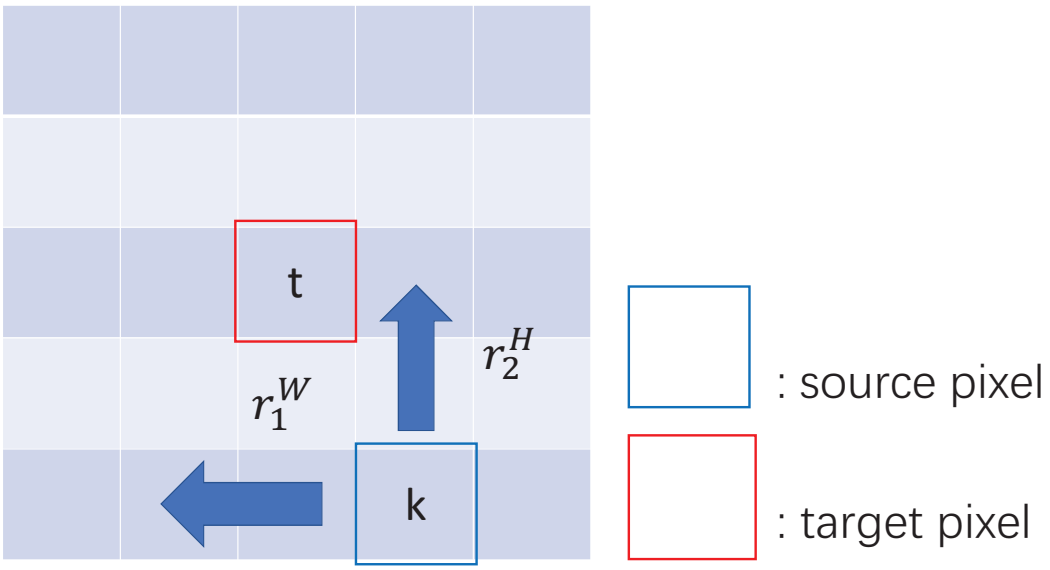

Figure 3. An example of Relative Position Weight calculation. Each grid represents a pixel of the feature map. The blue arrows show the orientations of the factorized relative position from the source pixel to the target pixel.

We apply a 3×3 convolution to the input feature map to obtain local feature aggregation and then concatenate the result to the PGA. Note that the channel of the local feature is also reduced to $c/2$ after aggregation. After concatenation on channels, the output of the PAM is the same as the input. Different from the convolution layers in ResNet50, all the convolutions in this module are not followed by BN and ReLU to ensure information independence and completeness in the integration process. The whole module is followed by a BN layer to normalize the output to reduce covariate shifts.

3.3. Loss Function and Optimization

We apply a cross-entropy loss [29] L_{ce} with a new label smoothing strategy and a triplet loss [30] L_{tri} for each feature vector yielding a global loss as

$$L = \lambda L_{ce} + (1 - \lambda) L_{tri} \tag{3}$$

where λ is a hyper-parameter ($\lambda = 0.6$ in all our experiments except the ablation study). The cross-entropy loss is defined as

$$L_{ce} = -\frac{1}{N} \sum_{i=1}^{N} q_i \log(p_i) \tag{4}$$

where N is the number of samples, p_i denotes the predicted probability for the ith identity, and q_i is the smoothed label, which is defined as

$$q_i = \begin{cases} \beta \times (1 - \epsilon) + \frac{\epsilon}{2N}, & \text{if } j = y_i \\ (1 - \beta) \times (1 - \epsilon) + \frac{\epsilon}{2N}, & \text{if } \|j - y_i\| = P \\ \frac{\epsilon}{2N}, & \text{others} \end{cases} \tag{5}$$

where j is the augmented image identity, y_i is the identity for the source image of the sample, P is the number of identities, ϵ is a precision parameter ($\epsilon = 0.1$ in all our experiments), and β is a symmetry factor ($\beta = 0.8$ in all our experiments except the ablation study). The Adaptive Label Smoothing makes the sequence of ranking similarities between feature embeddings from high to low: source images with the same IDs, source images and their mirrored images, source images and images with other IDs.

Each mini-batch samples p identities and n images per identity. A hard mining strategy is used to select the hardest positive and the hardest negative in each batch to form a triplet for calculating triplet loss as

$$L_{tri} = \sum_{i=1}^{p} \sum_{j=1}^{n} [m + \max_{k=1\cdots n} (D(f_{ij}, f_{ik}) - \min_{\substack{a=1\cdots p \\ c=1\cdots n \\ a\neq i}} D(f_{ij}, f_{ac})]_+ \tag{6}$$

where $D(\cdot, \cdot)$ denotes the Euclidean distances of two embeddings. Note that labels of mirrored images adopted in the triplet loss are the same as those of source images. This operation aims to force the distances of person image embeddings from different identities farther than those from self-mirrored images.

4. Experiments

4.1. Experimental Settings

Datasets. We evaluate MPGA-Net on four of the most widely used large-scale datasets, CUHK03-NP [31,32], Market1501 [33], DukeMTMC-ReID [34] and MSMT17 [35] all in standard protocols.

Training details. MPGA-Net is trained on four Nvidia GV100 GPUs. All images are resized into 384 × 128 pixels. Random erasing is adopted with a probability of 0.5 during training. In each mini-batch, we set $p = 32$ and $n = 8$. We employ Adam as the optimizer with a warm-up cosine annealing strategy for the learning rate [36]. The learning rate $Lr(t)$ is calculated as

$$Lr(t) = \begin{cases} 1.0 \times 10^{-3} \times \frac{t}{60}, & \text{if } t \leq 60 \\ 1.0 \times 10^{-3} \times \frac{1}{2}(1 + \cos(\pi \frac{t-60}{T-60}), & \text{if } 60 < t \leq T. \end{cases} \tag{7}$$

Evaluation details. All images are resized to 384 × 128 pixels and normalized. We follow the standard evaluation protocol in each dataset for a fair comparison and report Cumulative Matching Characteristics (CMCs) at Rank-1, 5, 10 and mean Average Precision (mAP) as evaluation metrics.

4.2. Comparison to State of the Art

We compare the performance of MPGA-Net with recent state-of-the-art person ReID methods on CUHK03-NP [31,32], Market1501 [33], DukeMTMC-reID [34], and MSMT17 [35] in Table 1, including methods based on ResNet [7,14,22,23,26,37–44], self-constructed network [6], neural architecture search [45], and tranformer [20,46–48]. Overall, our proposed MPGA-Net outperforms the state-of-the-art networks or achieves comparable performance. The mAP of our network on CUHK03-NP is less than that of C2F [42] mainly because the method uses an auxiliary-domain dataset for classification training while there are far fewer training images in CUHK03-NP than in other datasets. The performance of our method on MSMT17 is less than that of TransReID and DC-Former, which is mainly because of the base structure. Transformer has more parameters and high computational cost, so it has higher performance on large datasets. Meanwhile, we use an open-source post-processing method [32] for re-ranking to evaluate the feature effectiveness in mutual sample learning.

Table 1. Comparison with state-of-the-art person ReID methods.

Methods	CUHK03L		CUHK03D		Market1501		DukeReID		MSMT17	
	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP	Rank-1	mAP
PCB+RPP (2018) [14]	-	-	63.7	57.5	93.8	81.6	83.3	69.2	68.2	40.4
MHN (2019) [37]	77.2	72.4	71.7	76.5	95.1	85.0	89.1	77.2	-	-
OSNet (2019) [6]	-	-	72.3	67.8	94.8	84.9	88.6	73.5	78.7	52.9
ABDNet (2019) [26]	-	-	-	-	95.6	88.3	89.0	78.6	82.3	60.8
Pyramid (2019) [38]	78.9	76.9	78.9	74.8	95.7	88.2	89.0	79.0	-	-
IANet (2019) [7]	-	-	-	-	94.4	83.1	87.1	73.4	75.5	46.8
PISNet (2020) [39]	-	-	-	-	95.6	87.1	88.8	78.7	-	-
ISP (2020) [40]	76.5	74.1	75.2	71.4	95.3	88.6	89.6	80.0	-	-
RGA-SC (2020) [22]	81.1	77.4	79.6	74.5	**96.1**	88.4	-	-	80.3	57.5
SCSN (2020) [23]	86.8	84.0	84.7	81.0	95.7	88.5	91.0	79.0	83.8	58.5
CDNet (2021) [45]	-	-	-	-	95.1	86.0	88.6	76.8	78.9	54.7
PAT (2021) [41]	-	-	-	-	95.4	88.0	88.8	78.2	-	-
C2F (2021) [42]	80.6	79.3	81.3	**84.1**	94.8	87.7	87.4	74.9	-	-
DFLN (2023) [43]	86.8	84.0	84.8	81.5	95.9	89.8	91.3	81.8	-	-
SCS+ (2023) [44]	80.3	77.2	77.1	74.3	96.0	89.4	90.3	80.9	-	-
TransReID (2021) [20]	-	-	-	-	95.2	89.5	90.7	82.6	86.2	69.4
FED (2022) [46]	-	-	-	-	95.0	86.3	89.4	78.0	-	-
DCAL (2022) [47]	-	-	-	-	94.7	87.5	89.0	80.1	83.1	64.0
DC-Former (2023) [48]	84.4	83.3	79.6	77.5	96.0	90.6	-	-	**86.9**	**70.7**
MPGA-Net (Ours)	**88.0**	**85.7**	**85.8**	82.8	**96.1**	**90.9**	**93.1**	**84.4**	83.9	64.9
MPGA-Net + ReRanking [32]	91.8	92.8	90.8	91.1	96.5	95.6	94.5	92.4	86.3	77.9

4.3. Ablation Study

To demonstrate the effectiveness of the proposed data augmentation strategy, attention module, and label smoothing strategy on the performance of MPGA-Net, we incrementally evaluate each module on Market1501.

The impact of the proposed modules and strategies. Table 2 summarizes the experimental results of the ablation studies. The Baseline represents only the original backbone built on ResNet50. The performance of the Baseline decreases by using the HFNI strategy. It is mainly because images and their horizontal flipped images can hardly be distinguished without positions of salience parts or a balance of features and positions. The performance of the Baseline with PAM is similar to that without PAM, which indicates that it hampers the model to discover clues about the positions of salient parts in which mirrored images have the same labels as the source images. On the premise of the HFNI, each module or strategy brings effectiveness, which proves that relative position coordinated with visual chirality boosts the performance of MPGA-Net.

Table 2. Validity verification for each component within MPGA-Net on the Market1501 dataset.

Method	Rank-1	mAP
Baseline	94.8	85.6
Baseline + HFNI	91.2	80.5
Baseline + HFNI + ALS	95.2	88.1
Baseline + PAM	95.1	86.0
Baseline + PAM + HFNI + ALS	**96.1**	**90.9**

The impact of position for placing PAM. In Table 3, we conduct experiments to analyze the influences of PAM in different layers. The location of each layer is shown in Figure 2. As shown in Table 3, when selecting only one layer to insert the PAM in, the performances of Layer 2 and Layer 3 reach promising results. It proves that learning the relative positions of mid-level features is more effective. When inserting PAMs in

multiple layers, Layer 234 obtains the best performance. Adding PAMs to Layer 1 brings a performance decrease. It indicates that the relative position information about low-level features can hardly be utilized probably because of the disordered distribution in low-level features. In Table 4, we conduct experiments to analyze the influences of PAM in different blocks in Layer 234. 'last' represents the last residual block. 'odd' and 'even' represent the odd and even sequence of residual blocks respectively. Adding PAMs to the oven sequence set of blocks brings the best performance. By a comparative analysis of preset structures, we suppose that the process of Position-aware Global Attention would need a residual block following PAM. Moreover, inserting PAM to the first block in each layer may bring information loss because of the downsampling operation before the first block.

Table 3. The performances of different layers to place PAM on the Market1501 dataset.

Method	mAP	Rank-1	Rank-5	Rank-10
Baseline	85.6	94.8	98.2	99.0
Layer 1	86.5	94.6	98.5	99.0
Layer 2	90.5	95.7	**98.8**	**99.4**
Layer 3	90.5	**96.2**	98.7	99.2
Layer 4	87.9	95.2	98.4	99.0
Layer23	90.8	95.8	98.7	99.2
Layer34	90.7	95.8	98.6	99.2
Layer234	**90.9**	96.1	**98.8**	**99.4**
Layer1234	87.5	95.1	98.3	98.9

Table 4. The performances of different blocks to place PAM on the Market1501 dataset.

Location of PAM	mAP	Rank-1
last	87.2	94.3
odd	87.3	94.8
even	**90.9**	**96.1**

Parameter analysis of β. In Figure 4, we analyzed the effect of parameter β on the Adaptive Label Smoothing strategy. As source images with the same IDs have the same orientations of critical parts, their label logits should be larger than the mirrored images. The value range of β is from 0.5 to 1.0. We can observe that the performance of MPGA-Net reaches 90.3% mAP and 95.8% rank-1 with β set as 1.0, which means source images and mirrored images are forced to be separated. When β decreases, the performance improves and reaches 90.9% mAP and 96.1% rank-1 ($\beta = 0.8$), which indicates that source images and mirrored images have some identical feature patterns. With a further decrease of β, the performance decreases, which indicates that position information is of benefit to feature discrimination.

Parameter analysis of λ. In Figure 5, we analyzed the effect of parameter λ on the loss function formulation. The performance with λ greater than 0.5 is better than that with λ less than 0.5. It is mainly because the cross-entropy loss performs on new labels and the triplet loss performs on source old labels. It also confirms the significance of the orientations of relative positions in our design. Due to the mAP representing performance on all samples, we set λ to 0.6, with which the performance reaches the highest.

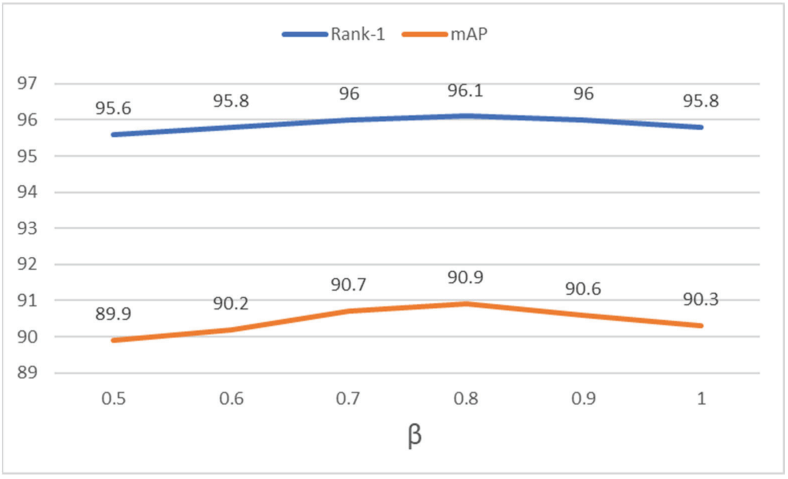

Figure 4. Parameter analysis of β on Market-1501.

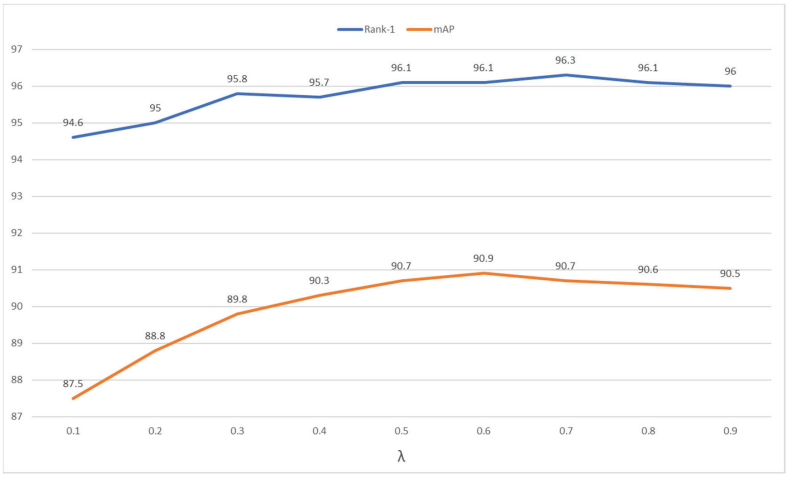

Figure 5. Parameter analysis of λ on Market-1501.

4.4. Visualization

Visualization of gradient responses. We apply the Grad-CAM [49] tool to our model for the qualitative analysis. Grad-CAM tool can indicate the regions that the network considers significant. Figure 6 shows the gradient responses of each layer in MPGA-Net. We can observe that each layer concentrates on different discriminative parts, especially the areas of vision chirality.

Visualization of matching results. We compare the ranking results of the baseline and MPGA-Net in Figure 7. As shown in Figure 7, MPGA-Net can effectively address the problem of similar clothing, self-occlusion, pose variations, and local omissions.

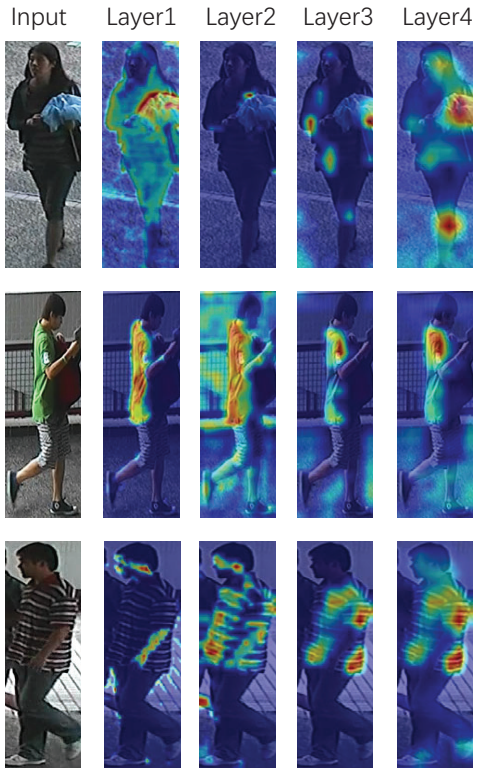

Figure 6. Grad-CAM [49] visualization for examples on each layer according to gradient responses. The original image is masked by a heat map with temperature representing the impacts. The impacts of the regions on results prediction from high to low are masked with colors: red, yellow, and blue.

Figure 7. Comparison of retrieval results on the Market-1501 dataset. (**a**) Baseline; (**b**) MPGA-Net. Figures with red boxes represent the wrong query results.

5. Conclusions

In this paper, we present a novel Multi-level Position-aware Global Attention Network to learn a global discriminative representation with an accumulation of relative position clues and attentioned local features. An Adaptive Label Smoothing strategy is proposed to balance representation learning and position mining. Extensive experiments on four challenging benchmarks have demonstrated that our proposed MPGA-Net achieves significant performance improvements.

However, our model needs more GPU memories and computation power than traditional CNNs, just like other self-attention and transformer-based methods. Moreover, due to the augmentation strategy, each batch should contain original images and mirrored images at the same time to train faster. On the large dataset MSMT17, the performance has greater room for progress. We will concentrate on how to extend our idea to new structures, such as transformers and large-convolution-based models.

Author Contributions: D.X., X.C. and J.C. conceived the experiments, D.X. conducted the experiments. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Nature Science Foundation of China (62071338, 62072347, U1903214, 61876135).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The publicly available datasets used in this research can be obtained through the following links: Market-1501: https://drive.google.com/file/d/0B8-rUzbwVRk0c054eEozWG9COHM/view (accessed on 26 February 2024), CUHK03: http://www.ee.cuhk.edu.hk/~xgwang/CUHK_identification.html (accessed on 26 February 2024), and MSMT17: https://www.pkuvmc.com (accessed on 26 February 2024).

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

Abbreviations

The following abbreviations are used in this manuscript:

ReID	re-identification
MPGA-Net	Multi-level Position-aware Global Attention Network
PAM	Position-aware Attention Module
PCB	Part-based Convolutional Baseline
HFNI	Horizontal Flipping with New Identities
ReLU	Rectified Linear Unit
ALS	Adaptive Label Smoothing
CBAM	Convolutional Block Attention Module
RPW	Relative Position Weight
PGA	Position-aware Global Attention
CMCs	Cumulative Matching Characteristics
mAP	mean Average Precision
BN	Batch Normalization

References

1. Zheng, L.; Yang, Y.; Hauptmann, A.G. Person Re-identification: Past, Present and Future. *arXiv* **2016**, arXiv:1610.02984.
2. Ye, M.; Shen, J.; Lin, G.; Xiang, T.; Shao, L.; Hoi, S.C. Deep Learning for Person Re-identification: A Survey and Outlook. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 2872–2893. [CrossRef] [PubMed]
3. 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.
4. Dai, Z.; Chen, M.; Gu, X.; Zhu, S.; Tan, P. Batch Dropblock Network for Person Re-identification and Beyond. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3691–3701.

5. 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; Volume 33, pp. 8295–8302.
6. Zhou, K.; Yang, Y.; Cavallaro, A.; Xiang, T. Omni-scale Feature Learning for Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3702–3712.
7. Hou, R.; Ma, B.; Chang, H.; Gu, X.; Shan, S.; Chen, X. Interaction-and-aggregation Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9317–9326.
8. 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.
9. Zhu, X.; Liu, J.; Wu, H.; Wang, M.; Zha, Z.J. ASTA-Net: Adaptive Spatio-temporal Attention Network for Person Re-identification in Videos. In Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1706–1715.
10. 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.
11. Li, W.; Zhang, Y.; Shi, W.; Coleman, S. A CAM-Guided Parameter-free Attention Network for Person Re-identification. *IEEE Signal Process. Lett.* **2022**, *29*, 1559–1563. [CrossRef]
12. Zhang, F.; Zhang, T.; Sun, R.; Huang, C.; Wei, J. An Efficient Axial-attention Network for Video-based Person Re-identification. *IEEE Signal Process. Lett.* **2022**, *29*, 1352–1356. [CrossRef]
13. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware Compositional 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. 2119–2128.
14. 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, Munich, Germany, 8–14 September 2018; pp. 480–496.
15. Sun, Y.; Xu, Q.; Li, Y.; Zhang, C.; Li, Y.; Wang, S.; Sun, J. Perceive Where to Focus: Learning Visibility-aware Part-level Features for Partial Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 393–402.
16. Lin, Z.; Sun, J.; Davis, A.; Snavely, N. Visual Chirality. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 1–19 June 2020; pp. 12295–12303.
17. 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.
18. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle Net: Person Re-identification with Human Body Region Guided Feature Decomposition and Fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1077–1085.
19. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned Bilinear Representations for Person Re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 402–419.
20. He, S.; Luo, H.; Wang, P.; Wang, F.; Li, H.; Jiang, W. Transreid: Transformer-based Object Re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 15013–15022.
21. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 6000–6010.
22. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware Global Attention for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 3186–3195.
23. Chen, X.; Fu, C.; Zhao, Y.; Zheng, F.; Song, J.; Ji, R.; Yang, Y. Salience-guided Cascaded Suppression Network for Person Re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3300–3310.
24. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
26. Chen, T.; Ding, S.; Xie, J.; Yuan, Y.; Chen, W.; Yang, Y.; Ren, Z.; Wang, Z. Abd-net: Attentive but Diverse Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 8351–8361.
27. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
28. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention Augmented Convolutional Networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3286–3295.

29. De Boer, P.T.; Kroese, D.P.; Mannor, S.; Rubinstein, R.Y. A Tutorial on the Cross-entropy Method. *Ann. Oper. Res.* **2005**, *134*, 19–67. [CrossRef]
30. Hermans, A.; Beyer, L.; Leibe, B. In Defense of the Triplet Loss for Person Re-identification. *arXiv* **2017**, arXiv:1703.07737.
31. 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.
32. Zhong, Z.; Zheng, L.; Cao, D.; Li, S. Re-ranking Person Re-identification with K-reciprocal Encoding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1318–1327.
33. 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.
34. Zheng, Z.; Zheng, L.; Yang, Y. Unlabeled Samples Generated by Gan Improve the Person Re-identification Baseline in Vitro. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3754–3762.
35. 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.
36. Loshchilov, I.; Hutter, F. Sgdr: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2016**, arXiv:1608.03983 .
37. Chen, B.; Deng, W.; Hu, J. Mixed High-order Attention Network for Person Re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 371–381.
38. Zheng, F.; Deng, C.; Sun, X.; Jiang, X.; Guo, X.; Yu, Z.; Huang, F.; Ji, R. Pyramidal Person Re-identification via Multi-loss Dynamic Training. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8514–8522.
39. Zhao, S.; Gao, C.; Zhang, J.; Cheng, H.; Han, C.; Jiang, X.; Guo, X.; Zheng, W.S.; Sang, N.; Sun, X. Do Not Disturb Me: Person Re-identification under the Interference of Other Pedestrians. In Proceedings of the European Conference on Computer Vision, Springer International Publishing: Cham, Switzerland, 23–28 August 2020; pp. 647–663.
40. Zhu, K.; Guo, H.; Liu, Z.; Tang, M.; Wang, J. Identity-guided Human Semantic Parsing for Person Re-identification. In Proceedings of the European Conference on Computer Vision, Cham, Switzerland, 23–28 August 2020; pp. 346–363.
41. 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.
42. Zhang, A.; Gao, Y.; Niu, Y.; Liu, W.; Zhou, Y. Coarse-to-fine Person Re-identification with Auxiliary-domain Classification and Second-order Information Bottleneck. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 598–607.
43. Yang, S.; Liu, W.; Yu, Y.; Hu, H.; Chen, D.; Su, T. Diverse Feature Learning Network With Attention Suppression and Part Level Background Suppression for Person Re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 283–297. [CrossRef]
44. Zhu, K.; Guo, H.; Liu, S.; Wang, J.; Tang, M. Learning Semantics-consistent Stripes with Self-refinement for Person Re-identification. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8531–8542. [CrossRef] [PubMed]
45. 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.
46. Wang, Z.; Zhu, F.; Tang, S.; Zhao, R.; He, L.; Song, J. Feature Erasing and Diffusion Network for Occluded Person Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4754–4763.
47. Zhu, H.; Ke, W.; Li, D.; Liu, J.; Tian, L.; Shan, Y. Dual Cross-attention Learning for Fine-grained Visual Categorization and Object Re-identification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 4692–4702.
48. Li, W.; Zou, C.; Wang, M.; Xu, F.; Zhao, J.; Zheng, R.; Cheng, Y.; Chu, W. DC-Former: Diverse and Compact Transformer for Person Re-Identification. *arXiv* **2023**, arXiv:2302.14335.
49. 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.