applied sciences

MDPI

Article

A Novel Dual Mixing Attention Network for UAV-Based Vehicle Re-Identification

Wenji Yin, Yueping Peng *, Zecong Ye and Wenchao Liu

PAP Engineering University, Xi'an 710086, China
* Correspondence: percy001@163.com

Abstract: Vehicle re-identification research under surveillance cameras has yielded impressive results. However, the challenge of unmanned aerial vehicle (UAV)-based vehicle re-identification (ReID) presents a high degree of flexibility, mainly due to complicated shooting angles, occlusions, low discrimination of top–down features, and significant changes in vehicle scales. To address this, we propose a novel dual mixing attention network (DMANet) to extract discriminative features robust to variations in viewpoint. Specifically, we first present a plug-and-play dual mixing attention module (DMAM) to capture pixel-level pairwise relationships and channel dependencies, where DMAM is composed of spatial mixing attention (SMA) and channel mixing attention (CMA). First, the original feature is divided according to the spatial and channel dimensions to obtain multiple subspaces. Then, a learnable weight is applied to capture the dependencies between local features in the mixture space. Finally, the features extracted from all subspaces are aggregated to promote their comprehensive feature interaction. In addition, DMAM can be easily plugged into any depth of the backbone network to improve vehicle recognition. The experimental results show that the proposed structure performs better than the representative method in the UAV-based vehicle ReID. Our code and models will be published publicly.

Keywords: dual mixing attention; UAV re-identification; deep learning

Citation: Yin, W.; Peng, Y.; Ye, Z.; Liu, W. A Novel Dual Mixing Attention Network for UAV-Based Vehicle Re-Identification. *Appl. Sci.* **2023**, *13*, 11651. https://doi.org/10.3390/app132111651

Academic Editors: Washington Yotto Ochieng, Wen-Long Shang, Kun Wang, Haoran Zhang and Yanyan Chen

Received: 13 September 2023
Revised: 6 October 2023
Accepted: 9 October 2023
Published: 25 October 2023

1. Introduction

Vehicle re-identification (ReID) [1–4] holds great importance in the realm of intelligent transportation systems (ITSs) in the context of smart cities. Vehicle ReID can be regarded as an image retrieval problem. Given a vehicle image, the similarity between each image and the image to be retrieved in the test set is calculated to determine whether the image to be retrieved is in the test set. Traditionally, license plate images have been employed for vehicle identification. However, obtaining clear license plate information can be challenging due to various external factors like obstructed license plates, obstacles, and image blurriness. Therefore, this work has been gradually abandoned by today's re-recognition methods. In addition, manually designing features is unsuitable for big data environments because it is only for specific tasks and focuses on unilateral features of images. Therefore, vehicle ReID based on vehicle appearance information will become the primary way to solve the above problems.

Thanks to the success of deep learning, the vehicle identification algorithm in the field of surveillance cameras again achieved impressive results [5–9]. According to the idea of solving the vehicle ReID problem, the methods of vehicle ReID can be divided into a global feature-based method, local feature-based method, attention mechanism-based method, vehicle perspective-based method, and generative adversarial network-based method. Typically, these methods [6,9–12] employ a deep metric learning model that relies on feature extraction networks. The objective is to train the model to distinguish between vehicles with the same ID and those with different IDs to accomplish vehicle ReID. However, as shown in Figure 1, there are discernible disparities between vehicle images

captured by UAV and those acquired through stationary cameras. The ReID challenge regarding UAV imagery introduces unique complexities stemming from intricate shooting angles, occlusions, limited discriminative power of top–down features, and substantial variations in vehicle scales.

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

Figure 1. Comparison of two types of vehicle images. There is a significant difference between the vehicle from the UAV perspective and the vehicle from the fixed camera. The vehicle under the fixed camera shooting angle is relatively fixed. In the view of UAVs, the shooting angles of cars are changeable, and there are many top–down shooting angles. (**a**) Surveillance cameras; (**b**) UAV cameras.

It is worth mentioning that traditional vehicle ReID methods, primarily designed for stationary cameras, face challenges in delivering optimal performance when adapted to the domain of UAV-based ReID. Firstly, the shooting angle of UAVs is complex. UAVs can shoot at different positions and angles, and the camera's viewpoint will change accordingly. This viewpoint change may cause the same object or scene to have different appearances and characteristics in different images. Second, the UAV can overlook or squint at a target or scene at different angles, resulting in viewpoint changes in the image. This viewpoint change may cause deformation or occlusion of the target shape, thus causing difficulties for feature extraction. To solve the above problems, it is necessary to add a mechanism [1,13–15] that can extract more detailed features when ReID extracts features to deal with the challenges brought by the drone perspective. The change in the UAV viewpoint makes the feature extraction algorithm need a certain robustness, which can correctly identify and describe the target in the case of significant changes in the viewpoint. The difference in the UAV view angle makes the feature extraction algorithm need to have the ability to adapt to shape changes and occlusions to improve the feature reliability and robustness in different views.

In recent years, the attention mechanism has gained significant popularity across multiple domains of deep convolutional neural networks. Its fundamental concept revolves around identifying the most crucial information for a given target task from a vast volume of available data. The attention mechanism selectively focuses on the image's different regions or feature channels to improve the model's attention and perception ability for crucial visual content. In the context of UAV-based vehicle ReID, the attention mechanism enables the model to enhance its perception capabilities by selectively highlighting the vehicle's specific regions or feature channels.

However, most attention mechanisms [15–18] focus on extracting features only from channels or spaces. The channel attention mechanism can effectively enhance essential channels, but it cannot deal with the problem of slight inter-class similarity. Spatial attention mechanisms can selectively amplify or suppress features in specific regions spatially, but they ignore the relationship between channels. To overcome the shortcomings of a

single attention mechanism, recent studies have begun to combine channel and spatial attention [19–21]. Such a hybrid attention mechanism can consider the relationship between channel and space at the same time to better capture the critical information in the input feature tensor. By introducing multiple branches of the attention mechanism or fusing different attention weights, the interaction between features can be modeled more comprehensively. Shuffle attention (SA) [19] divides molecular channels to extract key channel features and local spatial fusion features, with each subchannel acquiring channel and spatial fusion attention. The bottleneck attention module (BAM) [20] is a technique that generates an attention map through two distinct pathways: channel and spatial. On the other hand, the dual attention network (DANet) [21] incorporates two different types of attention modules on dilated fully convolutional networks (FCNs). These attention modules effectively capture semantic dependencies in both spatial and channel dimensions.

Most methods make the input feature map directly pass through the fused attention. At the same time, SA [19] can provide richer feature representation by dividing the subchannels, which better capture the structure and associations in images or other data. However, the SA [19] method of dividing the channel into subchannels mainly focuses on weighting the input features in the channel dimension, ignoring the possible details in the spatial dimension.

For that, our proposed DMAM, which combines SMA with CMA, in which the original feature is divided according to the dimensions of spatial and channel to obtain multiple subspaces. Each sub-feature map is processed independently, and the features of different channels and local regions can be extracted so that the network can better associate local features with the whole feature. Then, a learnable weight is applied to capture the dependencies between local features in the mixture space. In conclusion, the features extracted from multiple subspaces are merged to enhance their comprehensive interaction. This approach enables the extraction of more resilient features and leads to improved recognition accuracy. The key contributions of this method are outlined as follows:

- We introduce a novel DMANet designed to handle the challenges of unmanned aerial vehicle UAV-based vehicle ReID. DMANet effectively addresses issues related to shooting angles, occlusions, top–down features, and scale variations, resulting in enhanced viewpoint robust feature extraction.
- Our proposed DMAM employs SMA and CMA to capture pixel-level pairwise relationships and channel dependencies. This modular design fosters comprehensive feature interactions, improving discriminative feature extraction under varying viewpoints.
- The versatility of DMAM allows its seamless integration into existing backbone networks at varying depths, significantly enhancing vehicle discrimination performance. Our approach demonstrates superior performance through extensive experiments compared to representative methods in the UAV-based vehicle re-identification task, affirming its efficacy in challenging aerial scenarios.

The structure of the paper will be as follows: In Section 2, a comprehensive review and discussion of related studies is presented. The proposed approach is elaborated in Section 3, providing a detailed description. Following this, Section 4 presents the experimental results along with comparisons. Finally, conclusions are provided in Section 5.

2. Related Work

2.1. Vehicle Re-Identification

The ReID problem [1,22] is first explored and applied to humans. Compared with pedestrian ReID, vehicle ReID is more challenging. Firstly, vehicles tend to have high similarity in appearance, especially in the case of the same brand, model, or color. A higher similarity makes vehicle re-identification more challenging because relatively few features may distinguish different vehicles, and there is little difference between features. Second, vehicle re-identification may face more significant pose variation than human re-identification. Vehicles may appear at different angles, positions, and rotations, resulting in changes in the geometry and appearance characteristics of the vehicle, which increases

the difficulty of matching and alignment. Traditionally, vehicle Re-ID problems have been solved by combining sensor data with other clues [23–28], such as vehicle travel time [23] and wireless magnetic sensors [24]. Although the sensor technology can obtain better detection results, it cannot meet the needs of practical applications because of its high detection cost. Therefore, we should pay more attention to more cost-effective ways that can be viewed as basic. In theory, based on the vehicle license plate number, feature recognition technology is the most reliable and most accurate again [25,26]. However, the camera's multi-angle, illumination, and resolution significantly influence license plate identification accuracy. Additionally, criminals block, decorate, forge, or remove license plates, making re-identifying vehicles only by license plate information less reliable. Accordingly, researchers have considered vehicle attributes and appearance characteristics, such as shape, color, and texture [27,28].

With the development of neural networks, deep learning-based approaches have outshone others [5,29]. Significant changes in camera angles can lead to substantial differences in local critical areas for vehicle re-identification, which leads to low precision. The hybrid pyramidal graph network (HPGN) [30] proposes a novel pyramid graph network, targeting features closely connected behind the backbone network to explore multi-scale spatial structural features. Zheng et al. [5] proposed the deep feature representations jointly guided by the meaningful attributes, including camera views, vehicle types and colors (DF-CVTC), a unified depth convolution framework for the joint learning of depth feature representations guided by meaningful attributes, including camera view, vehicle type, and color of vehicle re-identification. Huang et al. [29] raised multi-granularity deep feature fusion with multiple granularity (DFFMG) methods or vehicle re-identification, which uses global and local feature fusion to segment vehicle images along two directions (i.e., vertical and horizontal), and integrates discriminant information of different granularity. Graph interactive transformer (GiT) [31] proposes a structure where charts and transformers constantly interact, enabling close collaboration between global and local features for vehicle re-identification. The efficient multiresolution network (EMRN) [32] proposes a multiresolution feature dimension uniform module to fix dimensional features from images of varying resolutions.

Although the current vehicle ReID method plays a specific role in the fixed camera perspective, the vehicle space photographed from the UAV perspective changes significantly, and extracting features from the top–down vertical angle is difficult. Moreover, the shooting angle of UAVs is complex. UAVs can overlook or squint at a target or scene at different angles, resulting in viewpoint changes in the image. The current method needs to be revised to solve the above problems well, and further research is required.

2.2. Attention Mechanism

The attention mechanism uses deep neural networks to imitate human cognitive processes. The method has been widely applied in computer vision, with the characteristic of learning more skills to express [8,9,33] Zhou et al. [33] constructed a motion attention transfer (MATNet) attention framework based on human visual attention behavior for semantic segmentation tasks, solving the problem of insufficient datasets of basic facts. In addition, a new weakly supervised semantic segmentation group learning framework [8] is proposed. The attention mechanism can also solve the problem of the detection and identification of human interaction in images (HOI) [9].

According to [34,35], for vehicle ReID, the attention mechanism concentrates on regions that correlate to delicate and distinct image areas, including windshield stickers and custom paints. The attention mechanism automatically extracts the characteristics of the distinct regions, increasing the vehicle re-identification task's accuracy. Khorramshahi et al. [34] found that most re-identification methods focus on the critical point locations. However, these point locations weigh differently in distinguishing cars. As a result, they created a dual-path adaptive attention model for the two-path vehicle ReID. The oriented conditional component appearance path learns to capture local discriminant

traits by concentrating on the most instructive critical spots, whereas the global appearance path catches macroscopic vehicle features. Teng et al. [35] presented a spatial and channel attention network based on diffusion-convolution neural network (DCNN). This attention model includes a spatial attention branch and a channel attention branch, which adjusts the output weights of different locations and channels separately to highlight the outputs in the distinguished regions and channels. The attention model refines the feature map and can automatically extract more discriminant features.

Although these attention mechanisms fuse attention to important areas or channels by combining different attention branches, mixed attention tends to learn along a single dimension, ignoring the features of the remaining dimensions. With significant space, perspective, and scene changes, the UAV perspective has excellent challenges for the attention mechanism.

3. Proposed Method

3.1. Dual Mixing Attention Module

We use a standard ResNet-50 as our backbone to extract features. Our proposed DMAM is shown in Figure 2. Because the vehicle space photographed from the UAV perspective changes significantly, and the shooting angle of UAVs is complex, UAVs can overlook or squint at a target or scene at different angles, resulting in viewpoint changes in the image. Currently, some attention mechanisms use channel or spatial attention mechanisms. However, spatial attention is more focused on the space region but ignores the different characteristics of the channel. Channel attention filters important feature channels while missing spatial features. Although partial attention mechanisms use channel and spatial fusion attention, mixed attention tends to learn along a single dimension, ignoring the remaining dimensions' features. To address this, we propose a novel DMAM to capture pixel-level pairwise relationships and channel dependencies, where DMAM comprises SMA and CMA. To enhance vehicle ReID, the DMAM can also be easily added to backbone networks at any level. We denote an input original feature as $\mathbf{X} \in \mathcal{R}^{C \times H \times W}$, which goes through DMAM and outputs enhanced feature as $\mathbf{Z} \in \mathcal{R}^{C \times H \times W}$. C denotes the number of channels, and H and W denote the height and width, respectively.

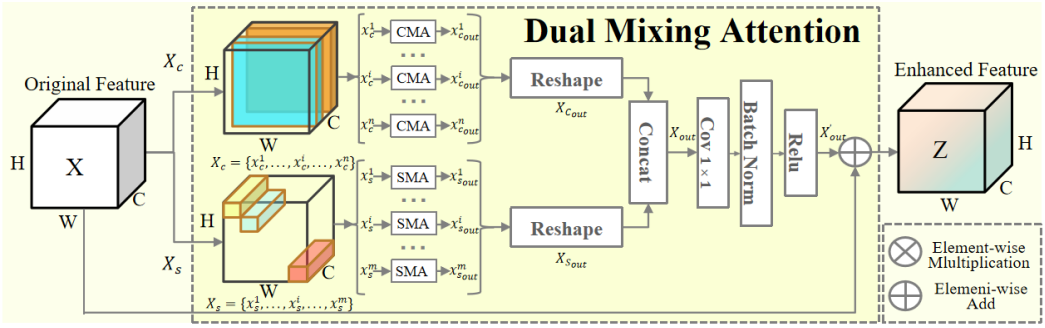

Figure 2. An overview of the proposed dual mixing attention module (DMAM) framework. The proposed method mainly contains four components: (1) The original feature is divided into two branches. (2) Feature map $\mathbf{X_c}$ divides n sub-channels $\mathbf{x_c^i}$ along the channel (C). Feature map $\mathbf{X_s}$ divides m subspaces $\mathbf{x_s^i}$ along space (HW). (3) Feature maps $\mathbf{x_c^i}$ and $\mathbf{x_s^i}$ were aggregated after entering channel mixing attention (cma) and spatial mixing attention (SMA), respectively. (4) After passing convolution (Cov), batch normalization (Bn), and rectified linear unit (Relu), the enhanced feature is finally output through the residual module.

Firstly, the original feature graph is split according to the dimensions of space and channel, and multiple subspaces are obtained, namely, $\mathbf{X_c} = \{\mathbf{x_c^1}, \mathbf{x_c^2}..., \mathbf{x_c^n}\}$ and $\mathbf{X_S} = \{\mathbf{x_s^1}, \mathbf{x_s^2}..., \mathbf{x_s^m}\}$. Each sub-feature map is processed independently, and the features of different channels and local regions can be extracted so that the network can better associate local features with the whole feature. Secondly, channel subfeature and spatial subfeature, $\mathbf{x_c^i}$ and $\mathbf{x_s^i}$, respectively, are sent into CMA and SMA to learn channels and spatial mixed features.

The output is $\mathbf{x}^{\mathbf{i}}_{\mathbf{c_{out}}}$ and $\mathbf{x}^{\mathbf{i}}_{\mathbf{s_{out}}}$. The dimensions of feature maps $\mathbf{x}^{\mathbf{i}}_{\mathbf{c}}$, $\mathbf{x}^{\mathbf{i}}_{\mathbf{c_{out}}}$ are $c/n \times h \times w$, and the dimensions of feature maps $\mathbf{x}^{\mathbf{i}}_{\mathbf{s}}$, $\mathbf{x}^{\mathbf{i}}_{\mathbf{s_{out}}}$ are $hw/m \times c$. The hybrid features increase the ability to input data model diversely, which better extracts complex feature representations from the UAV perspective. Third, the features extracted from their respective spaces are aggregated. After reshaping, the feature maps are transformed into $\mathbf{X_{c_{out}}}$ and $\mathbf{X_{s_{out}}}$. After concat, the output of the two feature maps is $\mathbf{X_{out}}$. The aggregation of features of multiple subspaces enhances the correlation between features. Integrating features promotes the interaction and information transfer between different features. Moreover, the generalization ability is also improved. Fourth, the feature map $\mathbf{X_{out}}$ passes through a set of 1×1 convolution (Cov), batch normalization (Bn), and rectified linear unit (Relu), with an output of feature map $\mathbf{X^{'}_{out}}$. The 1×1 Cov changes the dimension of the feature graph from $2C \times H \times W$ to $C \times H \times W$. Bn is normalized, changing the data distribution and preventing gradient explosion. As an activation function, Relu has low computational complexity, which improves the speed of the neural network gradient descent algorithm to better cope with significant changes in the vehicle size. Cov, Bn, and Relu effectively enhance the performance of the model and better learn complex or occluding features. Finally, the feature map $\mathbf{X^{'}_{out}}$ uses the residual structure to learn the original feature map through gap connections, which can accelerate the model's convergence rate, better use the information of previous levels, and better identify the features of the top–down vertical view shot by the UAV.

After several steps, we proposed a dual mixing attention module, including subspace segmentation, learning channels and spatial hybrid features, feature aggregation, convolutional normalization activation, and residual structure. DMAM can be connected behind the backbone network to make the features more profound and expressive, effectively coping with complicated shooting angles, occlusions, low discrimination of top–down features, and significant changes in vehicle scales. DMAM improves the robustness and discrimination of features, making the features more profound and expressive, thus improving the model's performance in ReID tasks.

3.2. Channel Mixing Attention

Channel mixing attention splits the feature map along the channel. Then, progressively merge the channel and spatial attention to obtain the CMA. The model can understand and represent the input data more comprehensively through this synthesis, improving features' distinguishing ability and generalization performance.

CMA is broken down into three phases as depicted in Figure 3: Firstly, the dimension of channel subfeature $\mathbf{x}^{\mathbf{i}}_{\mathbf{c}}$ is $C/N \times H \times W$, splitting along the channel into feature maps $\mathbf{x}^{\mathbf{i}}_{\mathbf{c_0}}$ and $\mathbf{x}^{\mathbf{i}}_{\mathbf{c_1}}$, as the input feature maps of space and channel attention. Secondly, the spatial input feature map $\mathbf{x}^{\mathbf{i}}_{\mathbf{c}}$ is multiplied by group norm (G_n) and shuffing (S_f) to extract the features of $\mathbf{x}^{\mathbf{i}}_{\mathbf{c}}$ space. More discriminative features are extracted by focusing on the critical space areas of vehicles through spatial attention. G_n calculates the mean and standard deviation, and divides the channels of each sample feature map into G groups. Each group will have C/G channels and then calculate the mean and standard deviation of the elements in these channels. Each group of channels is normalized independently with its corresponding normalization parameters. The formula for G_n is as follows:

$$G_n(X) = \frac{1}{\delta}(X - \mu), \tag{1}$$

where the parameter μ represents the mean, and the parameter δ illustrates the variance. The formula for μ is as follows:

$$\mu(x) = \frac{1}{(C/G)HW} \sum_{h=1}^{H} \sum_{w=1}^{W} X. \tag{2}$$

The formula for δ is as follows:

$$\delta(x) = \sqrt{\frac{1}{(C/G)HW}\sum_{h=1}^{H}\sum_{w=1}^{W}(X-\mu(x))^2 + \varepsilon}, \quad (3)$$

and the parameter ε is a tiny constant. Shuffing uses variable parameters to extract the feature weights of the subspace so that the model can better filter and focus on essential feature channels. The calculation formula is as follows:

$$S_f(x) = weight \times x + bias. \quad (4)$$

In the above equation, parameters *weight* and *bias* are variable parameters. By calculating the weight of each channel in the feature graph, it is possible to assign different importance to the features of different channels. The model's performance can be improved by focusing more intensively on the feature channels that are more helpful to the task.

In another branch, feature map $\mathbf{x_{c_1}^i}$ is sequentially subjected to adaptive average pool (A_p) and shuffing. The output is multiplied by feature map $\mathbf{x_c^i}$ to extract the spatial features of $\mathbf{x_c^i}$ feature maps. Spatial attention learns the importance of different locations and weight features according to this importance. A_p changes the dimensions of feature map $\mathbf{x_{c_1}^i}$ from $C/2N \times H \times W$ to $C/2N \times 1 \times 1$, which is used to extract the spatial features of the feature map.

Finally, the output dimension after concat is $C/N \times H \times W$. The output feature map $\mathbf{x_{c_{out}}^i}$ contains attention to feature graph channels and spaces. By mixing channel and spatial attention, the model can focus more precisely on different channels, locations, and correlations. The model can better adapt to different scales, shapes, and positions of vehicles photographed by UAVs.

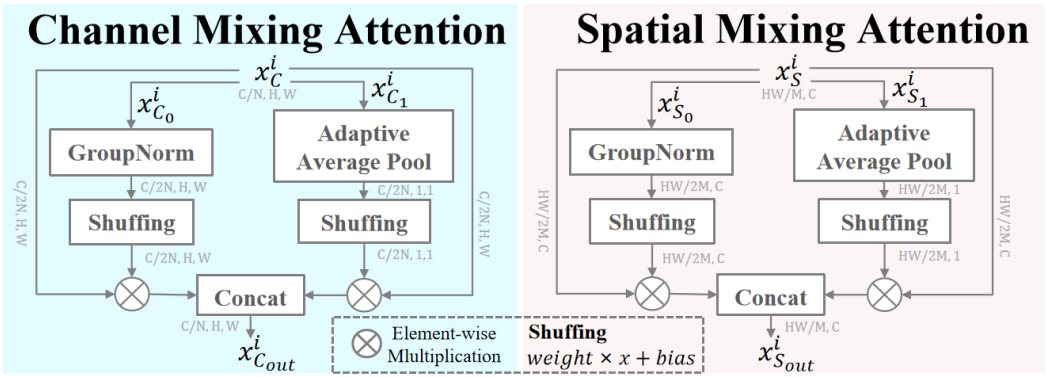

Figure 3. An overview of spatial mixing attention (SMA) and channel mixing attention (CMA). The process for CMA and SMA is similar. The CMA and SMA mainly contain three components: (1) The input feature map is divided into two sub-feature maps. (2) One branch is multiplied by GroupNorm and shuffing to extract the features of the space. After adaptive average pooling and shuffing, the other is multiplied by subfeature maps to extract channel features. (3) Finally, concat obtains the fusion feature.

3.3. Spatial Mixing Attention

SMA is similar to CMA. The difference between the two is that CMA groups feature maps in channel latitude, mainly focusing on the attention of each set of channels. In contrast, SMA groups feature maps in spatial dimensions. The input matrix $\mathbf{X_S} \in R^{HW/N \times C}$. In SMA, attention is used to weight and select features in the spatial dimension. However, in CMA, the attention mechanism is used to weight and choose features in the channel dimension.

The SMA was also divided into two branches, one of which went through G_n, shuffing, and multiplied with $\mathbf{x_s^i}$ to extract channel features. The other one went through the adaptive average pool,shuffing and multiplying by feature map $\mathbf{x_s^i}$ to extract spatial features. Finally, the output dimension after concat is $HW/M \times C$.

By dividing the molecular space, the model can divide the feature map into different subregions and calculate the attention weight of each subregion. This approach focuses more accurately on the importance of different spatial locations, allowing the model to capture local information about the vehicle better. This approach considers the relationship between different locations and channels while retaining the importance of spatial location. It helps the model understand the interaction of different channels at various locations, thus improving the consistency and accuracy of feature representation. It has great potential when dealing with complex UAV perspectives.

4. Analysis and Experiments

4.1. Datasets

UAV-VeID [36] dataset contains 41,917 images covering 4601 vehicles. The UAV-VeID dataset comprises videos taken by drones at locations such as highway interchanges, intersections, parking lots, etc., under different backgrounds and lighting conditions. The flight height of the captured drone is about 15 to 60 m, and the camera's vertical angle is between 40 and 80 degrees. This method of shooting causes the size and angle of view of the target vehicle to change.

VeRi-UAV [37] uses aerial photography to take pictures in multiple parking lots and on some roads. There is a training set and a test set for the VeRi-UAV. The test set has 4905 photographs of 111 automobiles, whereas the training set has 12,610 images of 343 vehicles. On this basis, about 15% of pictures are randomly selected from each vehicle ID to form a query set, and the remaining pictures form a gallery set.

Evaluation Metrics. We used the frequently used Top1, Top5, Top10, and mAP to assess how well various ReID techniques performed on datasets. They demonstrate how accurately the query sample matches the ID in the gallery. A high Top-k value indicates that query sample-based ID identification accuracy is high. mAP utilized in vehicle reidentification, is used to gauge how well the ReID approach performs overall. It displays the searchability of all test photos with the same ID. The search accuracy of the k position can be represented by the product of $P(k)G(K)$ and $P(k)G(K)$. In the search sequence, $G(k)$ denotes if the matched picture is present at position k, $P(k)$ denotes the likelihood that the first k search image contains the matched image, and $sum_k = 1nP(k)G(K)$ can denote the retrieval accuracy for query q as a whole. The AP of query q is the absolute retrieval precision divided by N_q. Formula (5) illustrates by defining mAP as the average of APs over all queries, and it may be used to evaluate the overall performance of a ReID model:

$$mAP = \frac{\sum_{i=1}^{M} AP(q_i)}{M}. \tag{5}$$

The evaluation metric $CMC@k$ is designed to describe the retrieval accuracy of matching locations as shown in Equation (6):

$$CMC@k = \frac{\sum_{i=1}^{M} F(q_i, k)}{M}. \tag{6}$$

If query qi's matched photos are in the top k images of the retrieved sequence, the evaluation index $F(q_i, k)$ indicates it. The evaluation metric $CMC@k$ measures the average search precision of all queries at position k in the search sequence. The most frequent $CMC@k$ values, Rank-1 and Rank-5, show the likelihood that an image will match in the top 1 and 5 positions of the retrieval sequence. This study uses mAP and $Rank\text{-}N$ as evaluation markers, similar to most of the ReID work.

4.2. Implementation Details

In this paper, we use the weight parameters of ResNet50 pre-trained on ImageNet as the initial weights of the network model. We follow [14,38] as follows: (1) All experiments were performed on PyTorch. Random identity sampling is taken for each training image

and resized to 256 × 256. (2) AMANet was trained for 150 epochs. (3) For the VeRi-UAV dataset, a stochastic gradient descent (SGD) optimizer was utilized with an initial learning rate 3×10^{-2}, and the training batch size was set to 32. The model received 60 epochs of training during this phase. The training batch size for the UAV-VeID dataset is 32, and the SGD optimizer's starting learning rate is 3×10^{-2}. (4) In addition, all test batch sizes are 64. (5) For the testing phase, our main assessment metrics are Rank-n and mean average precision (mAP).

4.3. Comparison with State of the Art

The performance comparisons of UAV-VeID and VeRi-UAV datasets between the previously related methods and our proposed DMANet are illustrated in Tables 1 and 2. As a whole, DMANet learning performs well compared with the others.

Table 1. Comparison of different methods on UAV-VeID (%).

Methods	Rank-1	Rank-5	Rank-10
Siamese-Visual [39]	25.98	41.98	50.61
VGG CNN M [40]	28.34	39.27	43.48
SCAN [35]	40.49	53.74	60.55
GoogleLeNet [41]	45.23	64.88	70.38
RAM [42]	45.26	59.35	64.07
CN-Nets [43]	55.91	76.54	82.46
TCRL [22]	56.44	77.21	82.98
EMRN [32]	63.47	79.84	84.66
CANet [1]	63.68	80.73	85.40
HPGN [30]	64.18	82.19	85.88
HSGNet [6]	64.22	85.31	86.36
AM + WTL [44]	69.11	87.23	91.64
GiT [31]	72.48	85.83	89.61
Baseline	70.94	84.56	88.22
Ours	**76.63**	**88.54**	**91.75**

4.3.1. Experiments on VeRi-UAV

On the VeRi-UAV dataset, the methods of comparison include [1,6,22,30–32,37,45–48]. Table 2 compares our proposed DMANet to other methods in the VeRi-UAV dataset. View decision-based compound match learning (VDCML) [37] is a method for vehicle ReID that learns the similarity between different views. Compared with DMANet, where Rank-1 and mAP are 29.7% and 32.4% higher than VDCML (ResNet50), although VDCML can extract effective features for vehicle ReID to a certain extent, it is more dependent on pre-defined rules or weight allocation, lacking automatic learning and adaptability. In contrast, the attention mechanism introduced by DMANet is more adaptive and expressive. DMANet achieves better results than VDCML in vehicle ReID tasks. To assess the importance of a feature based on all its components, contrastive attention net (CANet) [1] practices cooperation among the part features obtained by reweighting the part feature. Compared with DMANet, mAP is 9.1% higher than CAM. The reason is that DMANet can pay attention to the feature information of different scales to express the appearance characteristics of the vehicle more comprehensively and can capture a richer feature representation. In contrast, CAM is more dependent on the attention of limited-scale features and cannot fully use multi-scale feature information. Table 2 shows the comparison results with the methods mentioned above in detail. DMANet achieved excellent results. Compared with the baseline, DMANet shows an improvement of 7.45%, 1.17%, 1.52%, and 0.53% for different metrics of mAP, Rank-1, Rank-5, and Rank-10. In the end, we can come to the conclusion that DMANet comprehensively utilizes features of different levels and granularity to capture richer feature representations, achieving higher accuracy.

Table 2. Comparison of different methods on VeRi-UAV (%).

Methods	Rank-1	Rank-5	Rank-10	mAP
BOW-SIFT [45]	36.2	52.6	61.0	9.0
LOMO [46]	69.3	77.8	82.3	34.1
VGGNet [47]	56.0	72.4	78.6	44.4
ResNet50 [48]	58.7	74.0	79.5	47.3
VD-CML (VGGNet) [37]	62.5	76.2	81.3	49.7
VD-CML (ResNet50) [37]	67.3	78.8	83.0	54.6
TCRL [22]	77.1	79.2	84.9	58.5
EMRN [32]	87.6	88.9	92.4	65.9
CANet [1]	94.4	95.0	95.8	77.9
HPGN [30]	94.7	95.6	97.4	78.4
HSGNet [6]	94.8	95.7	97.6	78.5
GiT [31]	95.3	95.9	97.9	80.3
Baseline	95.1	95.6	97.5	79.6
Ours	**97.0**	**98.7**	**98.8**	**87.0**

4.3.2. Experiments on UAV-VeID

On the UAV-VeID dataset, the methods of comparison include [1,22,30–32,35,39–44]. Table 1 compares our proposed DMANet to other methods in the UAV-VeID dataset. For vehicle ReID, an efficient multiresolution network (EMRN) [32], which can implicitly learn collaborative multiresolution features via a unitary deep network, is proposed. Compared with DMANet, Rank-1 and Rank-5 are 13.16% and 8.7% higher than EMRN. DMANet can make better view decisions by mixing and integrating the feature information of different views. In contrast, EMN lacks fine processing and dynamic selection mechanisms in the view of decision making. DMANet achieves better results than EMRN in vehicle ReID tasks. In order to build a model for vehicle re-identification, graph interactive transformer (GiT) blocks are layered. In this model, graphs extract reliable global features, and transformers extract distinctive local features inside patches. Compared with DMANet, Rank-1 is 4.15% higher than GiT. DMANet can better model and handle changes in the vehicle's perspective, integrated channel and spatial feature information from different perspectives. However, it relies more on the transfer mode of the graph structure and cannot fully capture and utilize the critical features brought by the change of perspective. Table 1 shows that, compared with the baseline, DMANet contributes 5.69%, 3.98%, and 3.53% of the Rank-1, Rank-5, and Rank-10 improvement to the four subsets of UAV-VeID. In the end, by fusing attention mechanisms, DMANet can provide more flexibility so that the model can adaptively select and adjust different attention mechanisms according to the needs of specific tasks. This flexibility can help the model category and scale, better adapt to different perspectives such as diversity, and improve the generalization ability and adaptability of the model.

4.4. Ablation Experiment and Analysis

In this section, we design some ablation experiments to evaluate the effectiveness of our proposed DMANet, including (1) the role of the DMAM, (2) the effectiveness of which stage to plug the DMAM, (3) the effect of normalized strategy in DMAM, (4) the universality of different backbones, (5) comparison of different attention modules, and (6) the visualization of the model retrieval results.

4.4.1. The Role of Dual Mixing Attention Module

We evaluated the performance of different components of our proposed DMAM on the VeRi-UAV-based dataset in Table 3. The different results of using only CMA, SMA, and DMAM are listed separately. We made the following three observations:

(1) First of all, the results show that adding CMA to the baseline results in 1.20%, 1.62%, 0.55%, and 4.04% improvements in the assessment over the baseline on Rank-1,

Rank-5, Rank-10, and mAP, respectively. CMA introduces the attention mechanism, which can pay more attention to features with high distinction and importance to improve the feature judgment ability. Finally, the results show that CMA can indeed improve the model performance. (2) First, adding SMA to the baseline alone results in an improvement of 5.37% over the baseline on mAP. The addition of SMA improves the mAP by 1.33% compared with the addition of CMA. The difference is that SMA divides different subspaces along the space, and each subspace fuses after the attention mechanism. In vehicle ReID, there is a strong correlation between different vehicle parts. Each spatial position of the feature map can be weighted adaptively by the subspace through the attention mechanism so that the regional correlation between features can be better modeled and utilized. (3) First of all, adding DMAM, which combines CMA and SMA on top of the baseline, we can find another 1.9%, 3.02%, and 1.35%. Adding DMAM, which combines CMA and SMA on top of the baseline, we can find another 1.9%, 3.02%, 1.35%, and 7.40% improvement on Rank-1, Rank-5, Rank-10, and mAP, respectively. The addition of DMAM improves mAP by 2.03% compared with the addition of SMA. Both CMA and SMA have some limitations, and the fusion can complement each other's shortcomings and improve the robustness and generalization ability of the model. Ultimately, DMAM performs better than CMA and SMA.

Table 3. The role of dual mixing attention module (DMAM).

Methods	Rank-1	Rank-5	Rank-10	mAP
Baseline	95.14	95.63	97.48	79.59
+CMA	96.34	97.25	98.03	83.63
+SMA	96.56	97.42	98.27	84.96
Ours	**97.04**	**98.65**	**98.83**	**86.99**

4.4.2. The Effectiveness on Which Stage to Plug the Dual Mixing Attention Module

We designed a set of experiments and demonstrated its effectiveness by adding our proposed DMAM at different stages of the backbone network. Symbol ✓indicates that the DMAM is added after one of the residual blocks of the backbone network.

The experimental results of introducing DMAM after various backbone residual blocks are shown in Table 4. The findings demonstrate how the various residual blocks introduced to the backbone network impact the network's robustness. As shown in Figure 4, specifically, adding the DMAM behind the 4th (No. 2) and 5th (No. 3) residual blocks of the backbone improves the accuracy over the baseline (No. 0). The 5th (No. 3) residual block improves the accuracy over the baseline (No. 0), achieving 7.22% mAP, 1.17% Rank-1, 1.08% Rank-5, and 0.35% Rank-10 gains. It indicates that the 5th (No. 3) residual blocks of the module can effectively extract fine-grained vehicle features at these locations.

Table 4. The effectiveness on which stage to plug the dual mixing attention module (DMAM).

No.	Conv3_x	Conv4_x	Conv5_x	Rank-1	Rank-5	Rank-10	mAP
0				95.14	95.63	97.48	79.59
1	✓			95.74	96.88	97.74	82.95
2		✓		96.52	98.01	98.24	85.18
3			✓	**97.04**	**98.65**	**98.83**	**86.99**

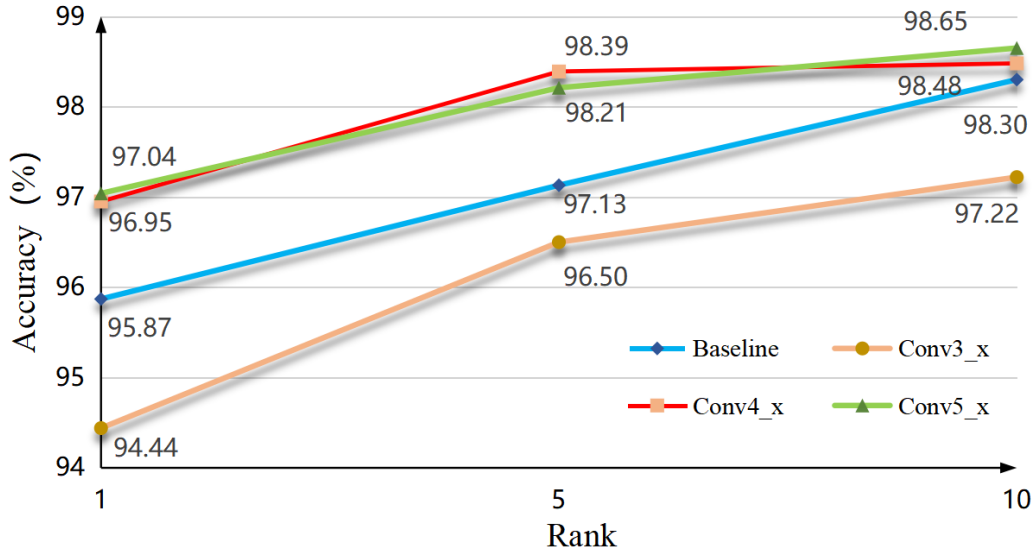

Figure 4. Ablation experiments with different backbone networks. Adding dual mixing attention module (DMAM) at different residual blocks of the backbone network on VeRi-UAV (%).

4.4.3. The Effect of Normalized Strategy in Dual Mixing Attention Module

To study the effect of the normalized strategy in our proposed DMAM, we replace the adaptive average pool (AAP) in DMAM with adaptive max pooling (AMP) or group norm (Gn) with instances norm (IN). Table 5 shows the ablation experiment performed on the VeRi-UAV dataset and draws three conclusions.

Table 5. The effect of normalized strategy in dual mixing attention module (DMAM).

Methods	Rank-1	Rank-5	Rank-10	mAP
Baseline	70.94	84.56	88.22	60.04
AAP → AMP	74.88	87.21	90.49	64.88
GN → IN	75.98	88.03	91.25	65.62
Ours	**76.63**	**88.54**	**91.75**	**66.22**

(1) First, the AMP normalization strategy is 4.84% higher than the baseline on the mAP. Secondly, AMP can enhance the representation ability of input features by selecting the most significant features for pooling. By highlighting the most discriminative and essential features, AMP helps extract critical information from vehicle images and strengthens the discrimination of vehicle appearance. Third, using AAP in DMAM improves mAP by 1.34% compared to AMP. Finally, AMP selects the most salient features in the input feature map for pooling, which means that other minor but still significant features are ignored. It may result in some key details and distinct loss features, reducing the vehicle image discriminant ability model.

(2) First, the IN normalization strategy is 5.58% higher than the baseline on the mAP. Secondly, IN is mainly used to normalize a single sample. The features of each sample distribution are more stable. In vehicle ReID, IN can be used to normalize the feature representation of each vehicle image to improve the network's performance. Third, using Gn in DMAM improves mAP by 0.6% compared to IN. Finally, the Gn characteristics of each group in normalization reduce the mutual influence between the characteristics of the channel. It makes the group norm changes for the batch size more robust in training and testing the model to maintain a stable performance. However, the batch size of IN is small, which may lead to a larger variance, introducing some instability.

(3) First, the DMANet is 6.18% higher than the mAP baseline. Secondly, the normalization strategy used in DMAM is shown in Figure 3. Gn is used to extract the spatial

features of the feature map, and AAP is used to extract the channel features of the feature map. In the image of a vehicle, regions at different locations have different textures, shapes, or detailed information, and Gn makes the network more robust in responding to different spatial locations. AAP can preserve key vehicle features and ignore unimportant details.

4.4.4. The Universality for Different Backbones

From the results of our proposed DMAM ablation in Figure 5, we observe the universality for different backbones. Firstly, the network performance will significantly improve when DMAM is added to different backbones. It can be seen that DMAM has strong adaptability. Second, the most significant improvement among them is Res2Net; the proposed DMANet can achieve 6.77% mAP and 5.87% Rank-1 gains on UAV-VeID. Finally, Res2Net enhances the network's receptive field and feature expression ability by introducing multi-scale feature representation. Res2Net can capture richer spatial information than the traditional single receptive field. In vehicle ReID tasks, the vehicle's appearance in different scales has rich details and shape characteristics, and the multi-scale features are helpful for heavy vehicle recognition tasks. It is suitable for complex tasks with vertical viewing angles and significant spatial changes, such as UAV viewing angle vehicle ReID.

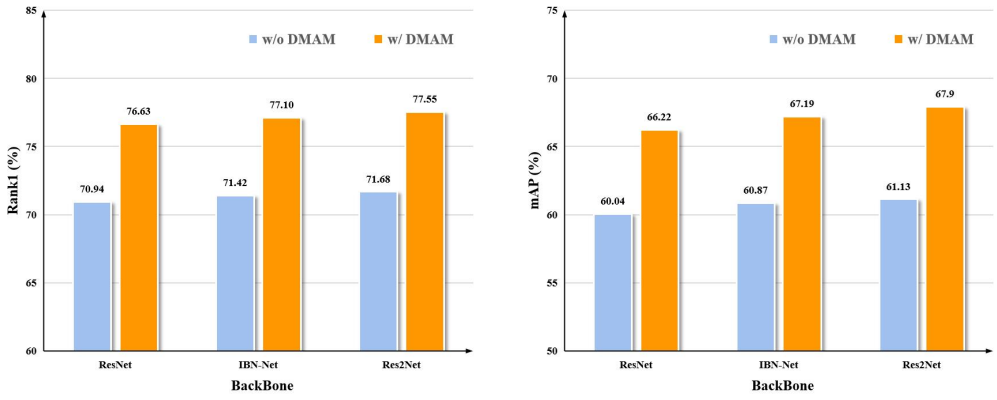

Figure 5. Comparison of dual mixing attention module (DMAM) in different backbones. The blue bar represents the performance of the original backbone, and the orange represents the performance after adding DMAM. (**a**) Rank1 results of different backbones; (**b**) mAP results of different backbones.

4.4.5. Comparison of Different Attention Modules

This subsection compares the performance with the already proposed attention modules [49–51]. Table 6 compares our proposed DMANet with different attentions on the VeRi-UAV dataset. Using a contrastive attention (CA) [1] module, we determine the significance of a feature based on the sum of all the parts. By reweighting the part feature, practical cooperation among the part features is derived. In comparison, for DMANet, Rank-1 and mAP are 2.6% and 9.12% higher than CA. DMANet allows for more equitable attention to the characteristics of different channels. It is essential for vehicle ReID tasks because the feature channels in different vehicle images may have different importance and expressiveness. Contrastive attention, by contrast, focuses only on the differences between positive and negative samples on each channel and may not be able to utilize the information from all channels fully. Polarized filtering and enhancement are two essential concepts for high-quality pixel-wise regression, and they are combined in the polarized self-attention (PSA) [51] block. In comparison, for DMANet, mAP is 6.29% higher than PSA. DMANet introduces more feature diversity by dividing multiple sub-feature maps into channel dimensions and spaces. Diversity can help improve the network's ability to capture the features in different vehicle images and improve the vehicle re-identification performance. However, PSA folds in both channel and spatial dimensions, which results in a certain level of information loss. It has a particular impact on the ability of fine-grained

feature discrimination in vehicle re-identification tasks. Table 6 shows that, compared with the baseline, DMANet contributes 6.29% of the mAP improvement of VeRi-UAV. In the end, by fusing attention mechanisms, DMANet can provide more flexibility so that the model can adaptively select and adjust different attention mechanisms according to the needs of specific tasks.

Table 6. Comparison of different attention modules.

Methods	Rank-1	Rank-5	Rank-10	mAP
CA [1]	94.44	95.02	95.83	77.87
SA [30]	94.72	95.57	97.43	78.42
SA&CA [6]	94.78	95.67	97.64	78.53
ACmix [49]	95.07	97.31	97.76	78.52
Cot [50]	95.87	97.31	97.76	80.30
Psa [51]	96.59	97.85	98.12	80.70
Ours	**97.04**	**98.65**	**98.83**	**86.99**

4.4.6. Visualization of Model Retrieval Results

To illustrate the superiority of our model more vividly, Figure 6 shows the visualization of the top 10 ranked retrieval results for the baseline and model on the VeRi-UAV dataset. Four query images corresponding to the retrieval results are randomly shown, the first row for the baseline method and the second for our method. The images with green borders represent the correct samples retrieved, while those with red edges are the incorrect ones retrieved.

Figure 6. Visualization of the ranking lists of model and baseline on VeRi-UAV. The top and bottom rows for each query show the ranking results for the baseline and joining the dual mixing attention module (DMAM), respectively. The green and red boxes denote the correct and wrong results, respectively.

As can be observed from Figure 6, the baseline usually focuses on the vehicle's appearance, such as color, shape, etc. Therefore, some negative matches appear due to pose and illumination similarity. Our proposed DMANet grouping features increases the variation and diversity of features, making distinguishing between vehicles with a similar appearance easier. In addition, DMANet fuses channels and spatial attention to enhance focus on critical areas and colors, shapes, and textures, making models distinguish similar-looking vehicles better. In summary, DMANet improves the model's ability to capture vehicle details, enhances vehicle differentiation, and thus improves the accuracy and robustness of vehicle reidentification.

5. Conclusions

In this paper, we proposed a novel DMANet to extract discriminative features robust to variations in viewpoint. Specifically, we first presented a plug-and-play DMAM, where DMAM is composed of SMA and CMA: First, the original feature was divided according to the dimensions of spatial and channel to obtain multiple subspaces. Then, a learnable weight was applied to capture the dependencies. Finally, the components extracted from all subspaces were aggregated to promote their comprehensive feature interaction. The experiments showed that the proposed structure performs better than the representative methods in the UAV-based vehicle ReID task.

Further Work. There are few datasets for vehicle ReID based on the UAV perspective, and the research space is ample. Consider extending the dataset regarding scenes such as car storage sites of large car factories or different lighting, and resolution. Furthermore, consider setting up datasets of changes in vehicle details, such as changes in the position of the vehicle decoration, the driver turning on the headlights, added passengers, and a sticker applied to the rear window. The model needs to determine whether the changed vehicle belongs to the same ID through the changes in details. This change aligns with reality and will present a significant challenge for vehicle ReID.

Author Contributions: Conceptualization, W.Y.; Methodology, W.Y.; Software, W.Y.; Validation, W.Y.; Formal analysis, Z.Y.; Resources, W.Y.; Data curation, Z.Y.; Writing—original draft, W.Y.; Writing—review & editing, Y.P. and W.L.; Visualization, W.L.; Project administration, Y.P.; Funding acquisition, Y.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by military equipment comprehensive research project [grant no. WJ20211A030131]; The PAP independently selected projects (grant no. ZZKY20223105); PAP Engineering University research innovation team project (grant no. KYTD201803).

Data Availability Statement: Data will be made available on request.

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

References

1. Li, M.; Wei, M.; He, X.; Shen, F. Enhancing Part Features via Contrastive Attention Module for Vehicle Re-identification. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; IEEE: Piscataway, NJ, USA, 2022 ; pp. 1816–1820.
2. Shen, F.; Peng, X.; Wang, L.; Zhang, X.; Shu, M.; Wang, Y. HSGM: A Hierarchical Similarity Graph Module for Object Re-identification. In Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 18–22 July 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–6.
3. Han, K.; Wang, Q.; Zhu, M.; Zhang, X. PVTReID: A Quick Person Re-Identification Based Pyramid Vision Transformer. *Appl. Sci.* **2023**, *13*, 9751. [CrossRef]
4. Qiao, W.; Ren, W.; Zhao, L. Vehicle re-identification in aerial imagery based on normalized virtual Softmax loss. *Appl. Sci.* **2022**, *12*, 4731. [CrossRef]
5. Li, H.; Lin, X.; Zheng, A.; Li, C.; Luo, B.; He, R.; Hussain, A. Attributes guided feature learning for vehicle re-identification. *IEEE Trans. Emerg. Top. Comput. Intell.* **2021**, *6*, 1211–1221. [CrossRef]
6. Shen, F.; Wei, M.; Ren, J. HSGNet: Object Re-identification with Hierarchical Similarity Graph Network. *arXiv* **2022**, arXiv:2211.05486.
7. Shen, F.; Shu, X.; Du, X.; Tang, J. Pedestrian-specific Bipartite-aware Similarity Learning for Text-based Person Retrieval. In Proceedings of the 31th ACM International Conference on Multimedia, Ottawa, ON, Canada, 29 October–3 November 2023.
8. Zhou, T.; Li, L.; Li, X.; Feng, C.M.; Li, J.; Shao, L. Group-wise learning for weakly supervised semantic segmentation. *IEEE Trans. Image Process.* **2021**, *31*, 799–811. [CrossRef] [PubMed]
9. Zhou, T.; Qi, S.; Wang, W.; Shen, J.; Zhu, S.C. Cascaded parsing of human-object interaction recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 2827–2840. [CrossRef]
10. Wu, H.; Shen, F.; Zhu, J.; Zeng, H.; Zhu, X.; Lei, Z. A sample-proxy dual triplet loss function for object re-identification. *IET Image Process.* **2022**, *16*, 3781–3789. [CrossRef]
11. Xie, Y.; Shen, F.; Zhu, J.; Zeng, H. Viewpoint robust knowledge distillation for accelerating vehicle re-identification. *EURASIP J. Adv. Signal Process.* **2021**, *2021*, 48. [CrossRef]
12. Xu, R.; Shen, F.; Wu, H.; Zhu, J.; Zeng, H. Dual modal meta metric learning for attribute-image person re-identification. In Proceedings of the 2021 IEEE International Conference on Networking, Sensing and Control (ICNSC), Xiamen, China, 3–5 December 2021; IEEE: Piscataway, NJ, USA, 2021; Volume 1, pp. 1–6.

13. Fu, X.; Shen, F.; Du, X.; Li, Z. Bag of Tricks for "Vision Meet Alage" Object Detection Challenge. In Proceedings of the 2022 6th International Conference on Universal Village (UV), Boston, MA, USA, 22–25 October 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 1–4.
14. Shen, F.; Wang, Z.; Wang, Z.; Fu, X.; Chen, J.; Du, X.; Tang, J. A Competitive Method for Dog Nose-print Re-identification. *arXiv* **2022**, arXiv:2205.15934.
15. Qiao, C.; Shen, F.; Wang, X.; Wang, R.; Cao, F.; Zhao, S.; Li, C. A Novel Multi-Frequency Coordinated Module for SAR Ship Detection. In Proceedings of the 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), Macao, China, 31 October–2 November 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 804–811.
16. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
17. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
18. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
19. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 2235–2239.
20. Park, J.; Woo, S.; Lee, J.Y.; Kweon, I.S. Bam: Bottleneck attention module. *arXiv* **2018**, arXiv:1807.06514.
21. 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.
22. Shen, F.; Du, X.; Zhang, L.; Tang, J. Triplet Contrastive Learning for Unsupervised Vehicle Re-identification. *arXiv* **2023**, arXiv:2301.09498.
23. Lin, W.H.; Tong, D. Vehicle re-identification with dynamic time windows for vehicle passage time estimation. *IEEE Trans. Intell. Transp. Syst.* **2011**, *12*, 1057–1063. [CrossRef]
24. Kwong, K.; Kavaler, R.; Rajagopal, R.; Varaiya, P. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors. *Transp. Res. Part C Emerg. Technol.* **2009**, *17*, 586–606. [CrossRef]
25. Silva, S.M.; Jung, C.R. License plate detection and recognition in unconstrained scenarios. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 580–596.
26. Watcharapinchai, N.; Rujikietgumjorn, S. Approximate license plate string matching for vehicle re-identification. In Proceedings of the 2017 14th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Lecce, Italy, 29 August–1 September 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6.
27. Feris, R.S.; Siddiquie, B.; Petterson, J.; Zhai, Y.; Datta, A.; Brown, L.M.; Pankanti, S. Large-scale vehicle detection, indexing, and search in urban surveillance videos. *IEEE Trans. Multimed.* **2011**, *14*, 28–42. [CrossRef]
28. Matei, B.C.; Sawhney, H.S.; Samarasekera, S. Vehicle tracking across nonoverlapping cameras using joint kinematic and appearance features. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 3465–3472.
29. Huang, P.; Huang, R.; Huang, J.; Yangchen, R.; He, Z.; Li, X.; Chen, J. Deep Feature Fusion with Multiple Granularity for Vehicle Re-identification. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 80–88.
30. Shen, F.; Zhu, J.; Zhu, X.; Xie, Y.; Huang, J. Exploring spatial significance via hybrid pyramidal graph network for vehicle re-identification. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 8793–8804.
31. Shen, F.; Xie, Y.; Zhu, J.; Zhu, X.; Zeng, H. Git: Graph interactive transformer for vehicle re-identification. *IEEE Trans. Image Process.* **2023**, *32*, 1039–1051. [CrossRef] [PubMed]
32. Shen, F.; Zhu, J.; Zhu, X.; Huang, J.; Zeng, H.; Lei, Z.; Cai, C. An Efficient Multiresolution Network for Vehicle Reidentification. *IEEE Internet Things J.* **2021**, *9*, 9049–9059. [CrossRef]
33. Zhou, T.; Li, J.; Wang, S.; Tao, R.; Shen, J. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Trans. Image Process.* **2020**, *29*, 8326–8338. [CrossRef]
34. Khorramshahi, P.; Kumar, A.; Peri, N.; Rambhatla, S.S.; Chen, J.C.; Chellappa, R. A dual-path model with adaptive attention for vehicle re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6132–6141.
35. Teng, S.; Liu, X.; Zhang, S.; Huang, Q. Scan: Spatial and channel attention network for vehicle re-identification. In Proceedings of the Advances in Multimedia Information Processing—PCM 2018: 19th Pacific-Rim Conference on Multimedia, Hefei, China, 21–22 September 2018; Proceedings, Part III 19; Springer: Berlin/Heidelberg, Germany, 2018; pp. 350–361.
36. Teng, S.; Zhang, S.; Huang, Q.; Sebe, N. Viewpoint and scale consistency reinforcement for UAV vehicle re-identification. *Int. J. Comput. Vis.* **2021**, *129*, 719–735. [CrossRef]
37. Song, Y.; Liu, C.; Zhang, W.; Nie, Z.; Chen, L. View-Decision Based Compound Match Learning for Vehicle Re-identification in UAV Surveillance. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–30 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 6594–6601.
38. Shen, F.; He, X.; Wei, M.; Xie, Y. A competitive method to vipriors object detection challenge. *arXiv* **2021**, arXiv:2104.09059.

39. Shen, Y.; Xiao, T.; Li, H.; Yi, S.; Wang, X. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1900–1909.
40. Chatfield, K.; Simonyan, K.; Vedaldi, A.; Zisserman, A. Return of the devil in the details: Delving deep into convolutional nets. *arXiv* **2014**, arXiv:1405.3531.
41. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
42. Liu, X.; Zhang, S.; Huang, Q.; Gao, W. Ram: A region-aware deep model for vehicle re-identification. In Proceedings of the 2018 IEEE International Conference on Multimedia and Expo (ICME), San Diego, CA, USA, 23–27 July 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
43. Yao, H.; Zhang, S.; Zhang, Y.; Li, J.; Tian, Q. One-shot fine-grained instance retrieval. In Proceedings of the 25th ACM International Conference on Multimedia, Mountain View, CA, USA, 23–27 October 2017; pp. 342–350.
44. Yao, A.; Huang, M.; Qi, J.; Zhong, P. Attention mask-based network with simple color annotation for UAV vehicle re-identification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
45. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a "siamese" time delay neural network. *Adv. Neural Inf. Process. Syst.* **1993**, *6*. [CrossRef]
46. 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.
47. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
48. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
49. 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.
50. Li, Y.; Yao, T.; Pan, Y.; Mei, T. Contextual transformer networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 1489–1500. [CrossRef]
51. Liu, H.; Liu, F.; Fan, X.; Huang, D. Polarized self-attention: Towards high-quality pixel-wise regression. *arXiv* **2021**, arXiv:2107.00782.