mathematics

MDPI

Article

Cross-View Multi-Scale Re-Identification Network in the Perspective of Ground Rotorcraft Unmanned Aerial Vehicle

Wenji Yin, Yueping Peng *, Hexiang Hao, Baixuan Han, Zecong Ye and Wenchao Liu

PAP Engineering University, Xi'an 710086, China; 20210058@ntit.edu.cn (W.Y.); hhx1214s@163.com (H.H.); hbx911wj@163.com (B.H.); yzc6666@yeah.net (Z.Y.); liuwch3@mail3.sysu.edu.cn (W.L.)
* Correspondence: percy001@163.com

Abstract: Traditional Re-Identification (Re-ID) schemes often rely on multiple cameras from the same perspective to search for targets. However, the collaboration between fixed cameras and unmanned aerial vehicles (UAVs) is gradually becoming a new trend in the surveillance field. Facing the significant perspective differences between fixed cameras and UAV cameras, the task of Re-ID is facing unprecedented challenges. In the setting of a single perspective, although significant advancements have been made in person Re-ID models, their performance markedly deteriorates when confronted with drastic viewpoint changes, such as transitions from aerial to ground-level perspectives. This degradation in performance is primarily attributed to the stark variations between viewpoints and the significant differences in subject posture and background across various perspectives. Existing methods focusing on learning local features have proven to be suboptimal in cross-perspective Re-ID tasks. The reason lies in the perspective distortion caused by the top-down viewpoint of drones, and the richer and more detailed texture information observed from a ground-level perspective, which leads to notable discrepancies in local features. To address this issue, the present study introduces a Multi-scale Across View Model (MAVM) that extracts features at various scales to generate a richer and more robust feature representation. Furthermore, we incorporate a Cross-View Alignment Module (AVAM) that fine-tunes the attention weights, optimizing the model's response to critical areas such as the silhouette, attire textures, and other key features. This enhancement ensures high recognition accuracy even when subjects change posture and lighting conditions. Extensive experiments conducted on the public dataset AG-ReID have demonstrated the superiority of our proposed method, which significantly outperforms existing state-of-the-art techniques.

Keywords: re-identification; Across Views; multi-scale network

MSC: 68T20

check for updates

Citation: Yin, W.; Peng, Y.; Hao, H.; Han, B.; Liu, W.; Ye, Z. Cross-View Multi-Scale Re-Identification Network in the Perspective of Ground Rotorcraft Unmanned Aerial Vehicle. *Mathematics* **2024**, *12*, 3739. https://doi.org/10.3390/math12233739

Academic Editor: Jonathan Blackledge

Received: 11 September 2024
Revised: 26 November 2024
Accepted: 26 November 2024
Published: 27 November 2024

1. Introduction

Re-ID is designed to locate the queried object across multiple cameras, with re-identifiable objects including people, vehicles, ships, animals, and so on. Traditional Re-ID schemes typically rely on multiple cameras from the same viewpoint to search for targets. Although these schemes exhibit differences in aspects such as background lighting between their training and test sets, they still do not fully simulate the complexity of the real world. To make Re-ID more aligned with practical application scenarios, researchers have begun to explore cross-domain Re-ID technologies where test data originate from different domains. The collaborative operation of fixed cameras and UAVs is gradually emerging as a new trend in the surveillance field. Faced with the significant viewpoint differences between fixed cameras and UAV cameras, the task of Re-ID faces unprecedented challenges.

At present, most of Re-ID's methods mainly rely on local feature learning. However, when a person is viewed from a drone's perspective, the visible features and the way they are presented are vastly different from those captured by a fixed camera at the ground level.

This is because the drone's view is often top-down, which provides a different point and captures different parts of the body or clothing. Next, the top-down view from a drone can capture features that are not as visible from a fixed camera's perspective. For example, the top of a person's head, the pattern on the back of their clothing, or the arrangement of objects they are carrying might be more discernible from above. Conversely, features that are prominent in a frontal or side view, such as facial features or the front of clothing, might be less visible or entirely obscured in a drone's view. Therefore, to address these challenges, Re-ID needs to learn the differences between various perspectives, including the drone perspective and the fixed camera perspective. This involves training models to understand and account for the changes in appearance that occur when the viewing angle changes. The learning process aims to capture a richer scale of features and details that are invariant to these perspective changes, and more discriminating.

To solve these problems, we propose a multi-scale scheme for ground-air cross-viewing angles that aims to address these challenges. To simulate a cross-perspective Re-ID training scheme, a cross-perspective multi-scale feature extraction network is proposed. By fusing multi-scale features, the impact of information loss is reduced, the network's ability is improved, and global and local information is utilized at different levels at the same time, so as to better understand the image content. In addition, to maintain the spatial consistency of features, we propose a cross-perspective alignment module to enhance the consistency of attention when the model faces different perspectives. Through the above scheme, we obtain a lightweight but efficient model architecture called multi-scale cross-view network.

Our contributions are as follows:

- We introduce a new type of MAVNet network designed to address the challenge of redefining characters across air and ground perspectives. MAVNet effectively solves problems related to large variations in viewing angles and scales, thereby enhancing robust feature extraction. Through a large number of experiments, our method shows better performance than the SOTA method in ground-air cross-perspective character recognition tasks, affirming its effectiveness in challenging cross-perspective tasks.
- Our proposed Multi-scale Across View Model (MAVM) uses a multi-scale convolution structure to learn features at different scales. This modular design facilitates comprehensive feature interaction, reduces loss of accuracy due to changes in viewing angles, and facilitates more differentiated feature extraction from different viewing angles.
- Our proposed AVAM optimizes the model's response to key parts by fine-tuning the attention weight, maintaining high recognition accuracy even under different heights of perspective.

2. Related Work

2.1. Multi-Scale Convolutional Networks

Early explorations of multi-scale convolutional networks focused on improving feature extraction capabilities by designing deeper or wider network structures. For example, VGGNet [1] captures richer layers of features by increasing network depth, while GoogleNet [2] introduces the Inception module to capture features at different scales in parallel. These networks have achieved remarkable success in image classification tasks. In recent years, an important development has been the introduction of the feature pyramid network (FPN) [3], which is particularly important in object detection and segmentation tasks. FPN [3] builds rich multi-scale feature representations through top-down paths and horizontal connections. In rerecognition tasks, image pixels differ greatly, so different levels of feature interaction are required. Res2Net [4] starts with the most basic common units in convolutional neural networks. The residual bottleneck structure widely existing in mainstream convolutional neural networks is enhanced in multi-scale. To solve the problem of image redundancy, OctaveConv [5] reduces the resolution of the low-frequency feature map; that is, the spatial dimension of the low-frequency feature map. This approach not only saves computing power and storage, but also helps each layer gain a larger receptive field to capture more contextual information. MixConv [6] is designed to design direct replacements for a single deep convolution, with the aim of easily taking advantage of different convolution kernel sizes without changing the network structure. They have achieved excellent performance in different visual tasks. These networks improve the model's ability to recognize objects at different scales by capturing and fusing features at different levels.

However, when current multi-scale networks are applied to cross-perspective Re-ID tasks, multi-scale networks may pay too much attention to details on small targets, while ignoring global context information, which is bad for cross-perspective generalization.

2.2. Different Perspectives on Re-ID Tasks

Re-ID is an important research direction in the field of computer vision, which aims to identify and track specific pedestrians from images or videos captured by different cameras. With the popularization of the monitoring system and the development of UAV technology, the research status of the Re-ID task can be summarized from three aspects: Re-ID of a fixed camera, Re-ID of UAV camera, and Re-ID task of cross-scene.

Fixed camera for Re-ID task. The Re-ID task of fixed cameras is one of the hot spots of research because they are widely deployed in urban surveillance, shopping malls, airports, and other situations. In this field, scholars mainly focus on how to improve the recognition accuracy of pedestrians and the robustness of the system. Typical work includes a deep learning-based pedestrian feature extraction method; for example, Mining Discriminative Features with Multi-Scale Contextual Attention (MSCAN) [7] proposed by Wang et al. MSCAN uses a multi-scale context attention mechanism to extract the distinguishing features of pedestrians. The main problems facing Re-ID's research include the influence of factors such as changes in viewing angle, lighting conditions, occlusion, and changes in pedestrian posture. In addition, the viewing angle limitations of fixed cameras also pose challenges for pedestrian tracking.

Drone camera for Re-ID task. The Re-ID task of drone cameras has been an emerging research direction in recent years. Compared to fixed cameras, drones provide more flexible viewing angles and wider coverage. However, this also brings new challenges, such as dynamic backgrounds, camera shake, and pedestrian size changes. Representative studies include Unsupervised Person Re-ID [8], which solves the Re-ID problem from the perspective of drones through unsupervised learning. The key problems that Re-ID's research needs to solve include how to adapt to the dynamic environment and improve the recognition accuracy of small target pedestrians.

Re-ID task across perspectives. The cross-view Re-ID task refers to the task of identifying the same pedestrian from different viewing angles, such as from a fixed camera to a drone camera. The challenge of this task lies in the differences in perspective, resolution, and background between different scenes. At present, the research on Re-ID in this aspect is relatively scarce. However, cross-perspective person datasets like AG-ReID have emerged, and our work will further promote the study of cross-perspective Re-ID.

3. Proposed Method

We proposed MAVM and the AVAM, two innovative components aimed at boosting cross-perspective Re-ID performance. The MAVM tackles the issue of angular variations in pedestrian imagery by integrating features across perspectives and managing network depth and width to avoid gradient problems, comprising four MAVs and two MAVDs for dimensionality reduction and feature enhancement. The AVAM aligns features amidst changes in viewpoint, lighting, and background, using an attention mechanism with query and key branches to target key features like silhouettes and textures. It includes three attention enhancement (AE) cycles and leverages convolutional operations to derive attention weights, which are adjusted for precise focus on critical features, maintaining accuracy under varying conditions.

To provide a comprehensive understanding of how these components synergize within our framework, we present the overall network architecture in Figure 1. This illustration delineates the structural composition of both the MAVM and AVAM, showcasing their integration and interaction within the broader system. The figure elucidates the flow of information and the functional interplay between these modules, which are pivotal to enhancing feature expressiveness and robustness.

3.1. Multi-Scale Across View Module

Our proposed MAVM, as depicted in Figure 1, is designed to address the challenge of angular variations in pedestrian imagery captured by different cameras, including UAV and fixed cameras. The significant angular changes can greatly alter the appearance and shape of pedestrians under different angles, complicating recognition tasks. To counteract this, MAVM is meticulously designed to balance network depth and width, avoiding issues of vanishing or exploding gradients.

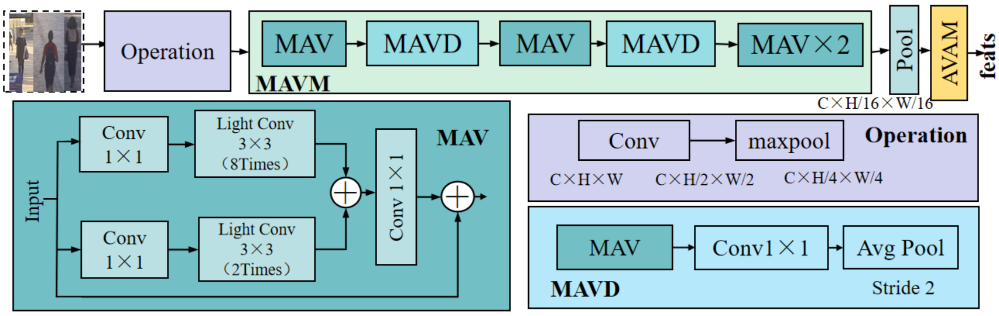

Figure 1. Multi-Scale Across View Module. The MAVM consists of four duplicate MAVs and two duplicate MAVD modules. The MAV is represented by the dark green area and the MAVD is represented by the blue area.

The MAVM is a sophisticated construct that integrates Multi-scale Across View (MAV) modules and Multi-scale Across View Downsampling (MAVD) modules. The input feature to the MAVM is denoted as $x \in R^{C\times H/4\times W/4}$, and the output feature is $x \in R^{C\times H/16\times W/16}$, where C represents the number of channels, and H and W represent the height and width of the feature maps, respectively. The MAVM consists of six modules, four duplicate MAVs, and two duplicate MAVD modules. The design choice of incorporating both MAVs and MAVDs is to ensure that our model can effectively capture and integrate features across different perspectives. The original feature maps are first entered into the MAV. After the two identical original feature maps undergo a 1×1 convolution, the two branches undergo two 3×3 convolutions and eight 3×3 convolutions, respectively, to extract features of different depths of the feature maps. Shallow feature maps usually contain more spatial details, while deep feature maps contain more abstract high-level semantic information. Secondly, the fusion of feature maps with different depths can form a richer and more comprehensive feature representation, which helps the model to better understand the appearance and attributes of pedestrians. Finally, a residual connection is added to the fused features after a 1×1 convolution to improve the utilization efficiency of the features. Following the MAV, a residual connection is introduced after a 1×1 convolution on the fused features. This design choice is intended to improve feature utilization efficiency. Residual connections are known to facilitate the training of deeper networks by mitigating the vanishing gradient problem and promoting the flow of information across the network, which is essential for feature expressiveness. The MAVD modules are introduced to downsample the feature maps, reducing spatial dimensions and computational load through 1×1 convolutions and average pooling. Average pooling is specifically chosen for its ability to improve spatial invariance, making the model less sensitive to minor input image changes such as translations and rotations. The output entering the MAVD for the first time is $x \in R^{C\times H/8\times W/8}$. After entering the same MAV and MAVD again, the output feature map is $x \in R^{C\times H/16\times W/16}$. The final output is obtained after entering the same two MAVs. The inclusion of residual connections after the 1×1

convolution on fused features is a deliberate design choice to enhance feature utilization efficiency. The MAVM, with its multi-scale fusion, residual structure, and downsampling makes the features more expressive, thereby improving the model's performance in cross-perspective Re-ID tasks.

3.2. Across View Alignment Model

Our proposed AVAM is shown in Figure 2. In the task of Re-ID, the images captured by different cameras have differences in angle of view, illumination, background, etc., and the change in posture during the movement of characters will affect the degree of matching between different images. To solve this problem, we propose AVAM to align features to increase their expressiveness. By inputting the original feature $x \in R^{C\times H\times W}$ into AVAM, we can obtain the output score as the similarity measure.

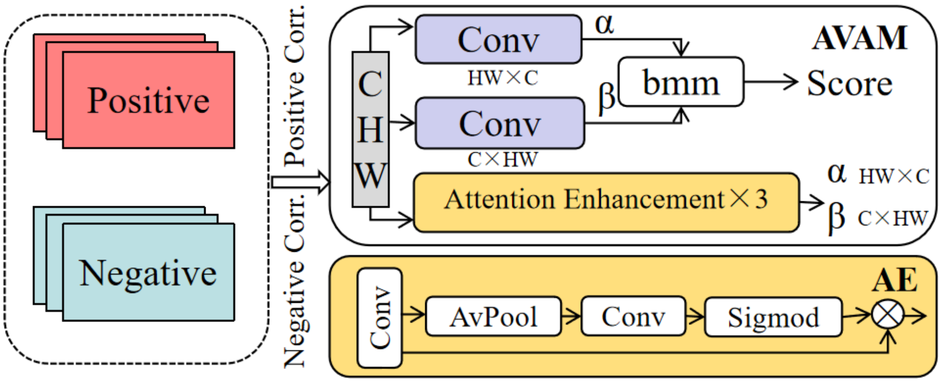

Figure 2. Across View Alignment Model. It mainly consists of Attention Enhancement and two branches: query and key. Score is the similarity measure of the output result obtained.

AVAM is composed of three repeated Attention Enhancement (AE) and query and key. First, AE is a residual structure composed of convolution, average pooling, and Sigmoid, which aims to enable the network to learn richer features to enhance feature representation. Too few cycles of AE will not have enough power to capture complex feature transformations, and too many cycles will lead to overfitting, making the model too complex. According to the Section 4.4.4 ablation experiment, three cycles is the best number. We will go through the AE method with $A_e(.)$. As an indication, the original feature passes through AE as follows:

$$P[A_e(x)] = \alpha, A_e(x) = \beta. \tag{1}$$

where $P[.]$ is permute, the obtained $\alpha \in R^{HW\times C}$, and the obtained $\beta \in R^{C\times HW}$.

Second, the two branches of query and key can be represented as follows: the terms "query" and "key" are integral to the attention mechanism that we employ to enhance feature representation. These components are derived from the original feature maps through a series of convolutional operations, as detailed below:

Query Branch: The query is generated by applying a permutation operation ($P[.]$) to the output of a convolution ($Conv(x)$). This operation rearranges the dimensions of the feature maps to facilitate the attention calculation. The query is represented as $Query = P[Conv(x)]$, where $Conv(x)$ denotes the convolution operation applied to the input feature x. The resulting query, $Query \in R^{HW\times C}$, is a matrix that will be used to compute the attention weights.

Key Branch: Similarly, the key is also derived from the convolutional output but is not subjected to the permutation operation. It is directly used as $Key = Conv(x)$. The key, $Key \in R^{C\times HW}$, is another matrix that, when combined with the query, allows the model to focus on the most relevant features for the task at hand.

$$Query = P[Conv(x)], Key = Conv(x). \tag{2}$$

where *Conv* represents the convolution operation. The final learned similarity measure will be expressed as follows:

$$Score = (\alpha Query) \times (\beta Key), \tag{3}$$

where × is expressed as matrix multiplication.

The query and key branches are derived from the original feature maps through a series of convolutional operations. The query is generated by applying a permutation operation to the output of a convolution, which rearranges the dimensions of the feature maps to facilitate the attention calculation. The key is derived directly from the convolutional output without permutation. This design allows the model to focus on the most relevant features for the task at hand, optimizing the model's response to key parts such as the silhouette of a person or the texture of clothing.

The final learned similarity measure is expressed as a matrix multiplication of the query and key. The attention mechanism adaptively adjusts to the salient features of individuals under varying imaging conditions, enhancing the robustness of local features while maintaining feature discrimination.

After several steps, our method can adaptively adjust the attention to the salient features of the people under different imaging conditions, thus enhancing the robustness of the local features while maintaining the feature discrimination. Specifically, AVAM optimizes the model's response to key parts (such as the silhouette of a person, the texture of clothing, etc.) by fine-tuning the attention weights, maintaining high recognition accuracy even when pose and illumination change.

4. Analysis and Experiments

4.1. Datasets

After an exhaustive review and comparison of available datasets, we have identified that the AG-ReID dataset stands out as the only one currently available that addresses cross-view Re-ID challenges. This highlights a significant gap in the field of cross-view Re-ID, where the scarcity of such datasets is a notable limitation. Our research underscores the importance of the AG-ReID dataset, as it provides a unique perspective in the domain of unmanned aerial vehicles (UAVs) and ground-based platforms. The comparative analysis of existing Re-ID datasets with AG-ReID, as detailed in Table 1, accentuates the distinctiveness of this dataset and underscores the need for further development in this area.

Table 1. Comparison of AG-ReID dataset with other publicly available person Re-ID datasets.

Datasets	Ground–Ground		Aerial–Aerial		Ground–Aerial
	Market 1501 [9]	**Duke MTMC** [10]	**PRAI 1581** [11]	**UAV Human** [12]	**AG ReID** [13]
IDs	1501	1404	1581	1144	388
Images	32,668	36,411	39,461	41,290	21,983
Views	fixed	fixed	mobile	mobile	fixed & mobile
Platforms	CCTV	CCTV	UAV	UAV	UAV & Phone
Altitude	<10 m	<10 m	20–60 m	2–8 m	15–45 m

We will modify the AG-ReID [13] dataset to accommodate the conditions set for cross-perspective requirements, thereby demonstrating and discussing the performance of cross-perspective Re-ID.

4.1.1. AG-ReID

The AG-ReID dataset was captured using a DJI XT 2 drone, which was flown at various altitudes ranging from 15 to 45 m to provide a diverse range of viewpoints and background contexts. Comprising 21,893 images across 388 unique identities, the dataset is divided into a training set featuring 199 identities with 11,554 images, and a testing set encompassing 189 identities with 12,464 images. The test set consists of 2033 query images and 10,429 gallery images, with the query images split into 1701 aerial and 962 ground-level images, complemented by 7204 aerial and 3255 ground-level gallery images. This arrangement facilitates the study of Re-ID across both ground and aerial vantage points.

4.1.2. Dataset Settings for Different Tests

In most Re-ID training Settings, the training set contains all the cameras, but this does not fit the reality, where the training and testing data sources are often separate. We separate the perspectives in the dataset, so that the perspectives in the training set and the test set in the separated data set will no longer coincide, so as to test the performance of the algorithm for cross-perspective tasks. We have innovatively processed the dataset as follows:

Train the ground and test query the data set AG1 in the air. This setting will remove the drone view from the training set, leaving only the fixed camera view, while query and gallery will both be set to images taken from the drone view. The specific dataset configuration is shown in Table 2.

Table 2. The number of scenes and Train images, Query images, and Gallery images in the AG1 and AG2 datasets.

Datasets	Scenarios	Train Images	Query Images	Gallery Images
AG 1	Train_Ground	3400	1071	7204
AG 2	Train_Aerial	8154	962	3255

Train the air, test query the ground dataset AG2. This setting will remove the fixed camera view from the training set, leaving only the drone camera view, while query and gallery will both be set to images taken from the fixed camera, as shown in Figure 3.

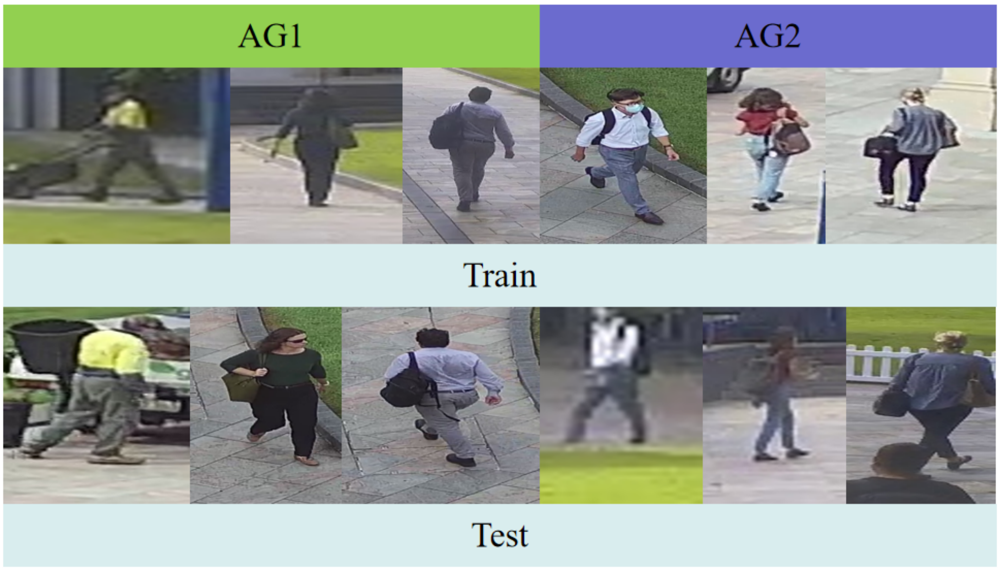

Figure 3. Legend for datasets AG1 and AG2.

Following the changes, we crafted a cross-view dataset that distinctly categorizes ground-level and aerial perspectives. The composition of this dataset, including the count of images for each category, is detailed in Table 2. This curated collection is specifically engineered to evaluate the proficiency of our proposed algorithm in handling substantial angular variations, thereby offering a more authentic testbed for its performance.

4.1.3. Evaluation Metrics

To comprehensively evaluate the performance of the proposed methods, two generally accepted evaluation indicators were adopted in this study: mean Average Precision (mAP) and Cumulative Matching Characteristics (CMC-k), or Rank-k matching accuracy. The mAP index is calculated by averaging the search performance across multiple real tags, providing a comprehensive evaluation of the overall performance of the search system. CMC-k, on the other hand, focuses on the probability of identifying the correct match in the first k search results, which provides a probability measure to evaluate the recognition ability of the system in the first k rankings. Specifically, the Rank-1 matching accuracy reported in this study is that for CMC-k, and we set k = 1, which reflects the ability of the system to identify the correct match in the first search result. This kind of evaluation is critical to understanding the performance of the system on the most likely matches and is especially critical for the evaluator to re-identify the system. Through these two indices, we can comprehensively understand the accuracy and reliability of the proposed method in the task of target recognition.

4.2. Implementation Details

We performed our experimental evaluation utilizing AG1 and AG2 datasets. The model parameters were optimized using stochastic gradient descent (SGD) with an initial learning rate set to 0.025, over the course of 350 epochs. A warm-up phase was implemented for the initial 10 epochs, after which the learning rate was decremented by 0.1 at the 150th, 225th, and 300th epochs. For the refinement of architectural parameters, the Adam optimizer [14] was engaged, and initialized with a learning rate of 0.002. The images were processed at a resolution of 128 by 256 pixels.

4.3. Comparison with State of the Art

Tables 3 and 4 give a comparison of the performance of the AG1 and AG2 datasets with our proposed MAVNet. Overall, MAVNet performed better than other methods.

Table 3. Comparison of different methods on AG1(%).

Models	mAP	Rank1
CAL [15]	12.0	53.5
MSINet [16]	50.4	90.6
Strong-baseline [17]	52.3	95.8
Generalizing [18]	54.2	96.0
Ours	58.2	96.9

Table 4. Comparison of different methods on AG2(%).

Models	mAP	Rank1
CAL [15]	7.4	30.4
Strong-baseline [17]	33.9	80.8
MSINet [16]	34.1	81.1
Generalizing [18]	35.2	82.1
Ours	45.8	85.2

4.3.1. Experiments on AG1

On the AG1 dataset, the comparison methods include [15,17,18]. Table 3 compares our proposed MAVNet with other methods for conducting trials in AG1 datasets. Causal inference-based counterfactual attention learning (CAL) [15] analyzes the effects of learned visual attention on network prediction through a counterfactual intervention to encourage fine-grained image recognition for network learning attention. Cross-perspective Re-ID involves significant changes in characters from different perspectives, and it is necessary to capture key features from different perspectives. CAL [15] quantifies the quality of attention by comparing the effect of facts (learned attention) and counterfacts (uncorrected attention) on the final prediction. When the learned attention is too focused on some local features, it may not cover the comprehensive features required for cross-perspective Re-ID, resulting in poor performance in cross-perspective datasets. In the cross-view Re-ID task, Strong-baseline achieved better results than CAL. Strong-baseline proposes a novel neck structure named batch normalization neck (BNNeck). Strong-baseline [17] uses tricks to improve Re-ID's ability model and only use the global features extracted by the model. Compared with CAL, BNNeck helps to solve the inconsistency of measurement loss and classification loss in the same embedded space and helps to learn more differentiated feature representations, which is especially important for cross-perspective Re-ID. Compared to CAL [15], MAVNet introduced AVAM, which helps the model maintain consistency of attention when faced with images from different perspectives, better learn distinguishing features from limited data, and reduce misjudgments due to changes in perspective. In this section, we conducted a series of ablation studies to assess the individual contributions and overall performance of our proposed MAVNet framework. The results, as delineated in Table 3, demonstrate the robust performance of MAVNet. Enhancements over the generalizing model are evident, with improvements of 4.0% in mAP. These gains underscore MAVNet's adept integration of multi-level and multi-granularity features, culminating in more comprehensive feature representations and superior accuracy.

4.3.2. Experiments on AG2

On the AG2 dataset, the comparison methods include [15,17,18]. Table 4 compares our proposed MAVNet with other methods in the AG2 dataset. Generalizing ReID [18] proposed a novel cross-domain mixup scheme. It alleviates the abrupt transfer by introducing the interpolation between the two domains as a transition state. Compared to CAL [15] and Strong-baseline, Generalizing ReID performed better, 1.9% higher on mAP, because Generalizing ReID [18] imposed constraints on matching under the same camera and matching under different cameras. To reduce the retrieval bias caused by camera differences, improve the ability of cross-view retrieval. Table 4 shows that, compared to MAVNet, mAP is 10.6% taller than Generalizing ReID. This is because MAVNet can focus on feature information at different scales, and can capture richer feature representations, which is better for Re-ID with large differences in perspective. In contrast, Generalizing ReID relies more on focusing on finite-scale features and cannot make full use of multi-scale feature information. Through the multi-scale feature network, MAVNet can learn a wider range of features, allowing the model to focus on more discriminative features from the aerial view of the drone and the fixed camera view. This flexibility can help the model adapt to different camera angles and improve the generalization ability of the model.

4.4. Ablation Experiment and Analysis

In this section, we conducted a series of ablation studies aimed at assessing the efficacy of the proposed MAVNet. The experiments were structured to evaluate various components of our model, including: (1) the Role of Multi-scale Across View Module and Across View Alignment Module, (2) The Effectiveness on Multi-scale Across View, (3) The Effectiveness on Different Scales in Multi-scale Across View, (4) The Effectiveness on Different Depths Across View Alignment Model, and (5) Visualization of Model Retrieval Results.

4.4.1. The Role of Multi-Scale Across View Module and Across View Alignment Module

In Figure 4, we present an assessment of the efficacy of various components within our MAVNet architecture when applied to a dataset derived from AG1. This figure provides a comparative analysis of the performance outcomes utilizing solely the AVAM module, the MAVM module, and the combined effect of both, as well as the baseline structure which does not include either the AVAM or MAVM modules. Our analysis led to the following key observations:

A. First, the results show that, after adding AVAM to the baseline, mAP, and Rank1 results improved by 0.8% and 0.4% compared to the baseline. AVAM is able to improve the accuracy of recognition by explicitly aligning spatial attention between images from across perspectives, helping models correctly and consistently focus on specific parts of differentiated people from different perspectives. The results show that AVAM can indeed improve the performance of the model.

B. Second, by adding MAVM alone to the baseline, the improvement for mAP was of 1.2% and for Rank1 it was of 1.8% over the baseline. In the AG1 dataset, the main training is the character pictures from the fixed camera perspective, and the character features are rich and extensive compared with the UAV perspective. Multi-scale networks can generate more robust feature representations, which helps to identify and distinguish objects from different viewing angles.

C. Third, by adding the combination of MAVM and AVAM to the baseline, we can find improvements of 2.4% and 2.9% for Rank-1 and MAP, respectively. MAVM can accommodate images with different resolutions and viewing angles, which is ideal for processing images taken from the air and the ground. Changes in perspective can cause significant changes in the appearance of an object, and multi-scale networks can reduce this effect by capturing multi-scale features. Alignment helps to maintain the feature differentiation in multi-scale feature fusion and enables the model to adapt to different data distributions, thus improving the retrieval accuracy in the retrieval task. Thus, MAVNet outperforms MAVM and AVAM alone.

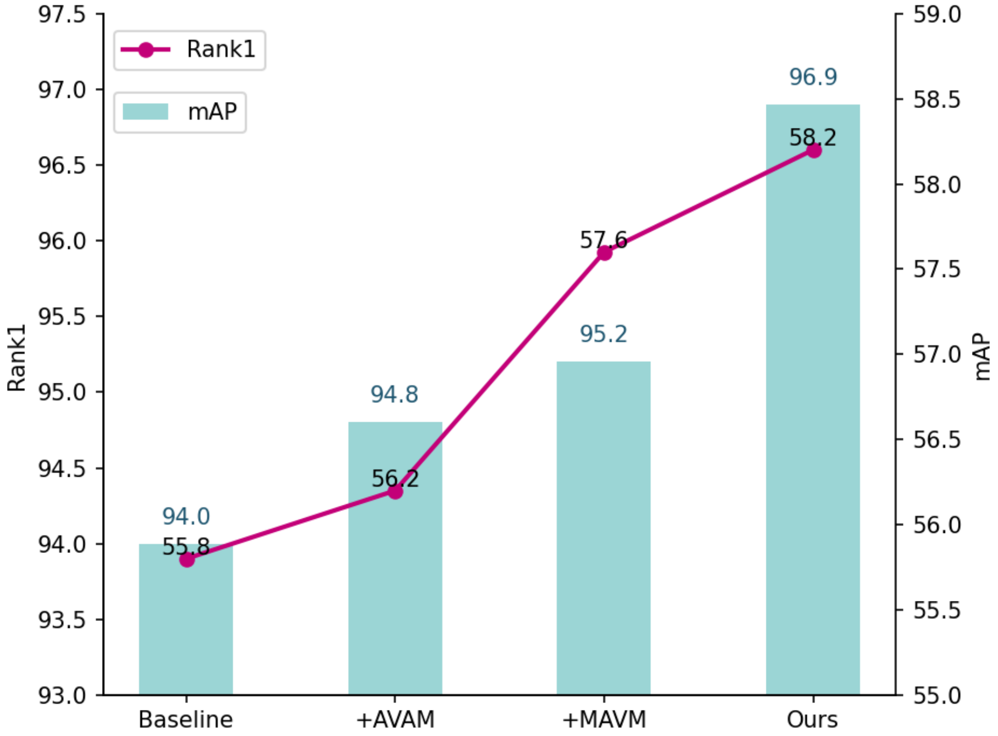

Figure 4. The performance of different modules of MAVNet on AG1-based datasets, with blue representing mAP and red representing Rank1.

4.4.2. The Effectiveness on Multi-Scale Across View

As shown in Figure 5, the MAV is divided into two Multi-Scale Models and two Operation regions. We designed a set of experiments to explore the optimal setting of the MAV at each stage by discussing different strategies in the Operation area of the MAV. There are six MAVs in the entire MAVNet for a total of 12 operations. The red arrow in Figure 5 indicates that no operation is used, and the green arrow indicates that the features of the two branches are exchanged at this stage.

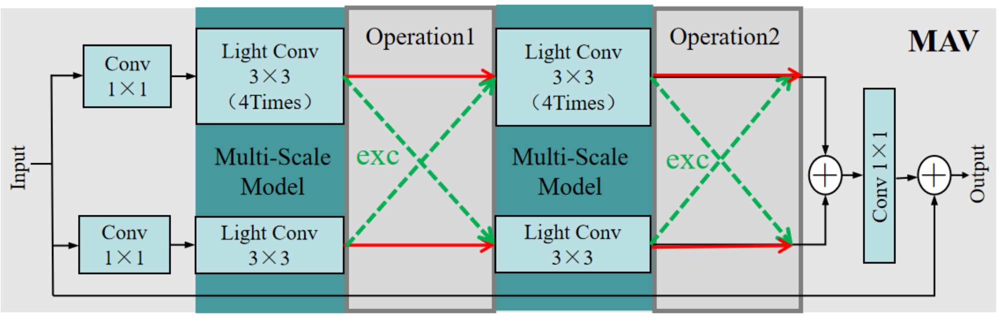

Figure 5. In the specific structure diagram of the MAV, the dark green area is defined as the Multi-Scale Model, and the area framed by gray is defined as the operation.

In the experiments of whether features were exchanged in 12 operations, Rank1 and mAP obtained by different operations are shown in Figure 6. These results demonstrate the effectiveness of the MAV setup. Specifically, Rank1 and mAP, obtained by analyzing 12 operations without exchanging features according to the bar chart, are the highest, and far higher than the MAV of any exchanging features. In the AG1 dataset, the character pose changes, the noise is large, and the resolution is low. Multi-scale feature exchange will make the consistency between features not high, which may cause the model to learn the wrong association, which will reduce the performance. Feature swapping can make the model too sensitive to specific details in the training data, increasing the risk of overfitting and being unsuitable for the task at hand. The results show that the performance of MAV is better without feature exchange.

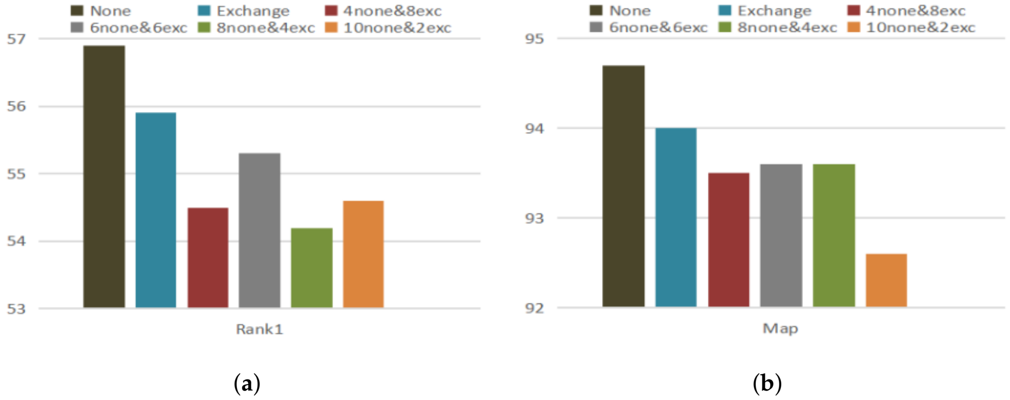

Figure 6. (**a**) Rank1 from different operations. (**b**) mAP from different operations.

4.4.3. The Effectiveness on Different Scales in Multi-Scale Across View

As shown in Figure 5, the MAV is divided into two Multi-Scale Models and two Operation regions. We designed a set of experiments to explore the optimal setting of the MAV by discussing the different scales used in the Multi-Scale Model area highlighted in dark green in the figure.

The mAP and Rank1 of different scales in the Multi-Scale Model area are shown in Table 5. Conv1 & 2 indicates that the two branches of MSM in the figure use one convolution and two convolution respectively, and Conv1 & 2 & 3 indicates that the MSM will be divided into three branches using 1, 2, and 3 convolutions, respectively. The effect of MSM using two branches is generally better than that of using three branches. First of all, the two-branch model is sufficient to meet the requirements of the task of cross-ground and air-view Re-ID, while the three-branch model may lead to overfitting due to too many parameters. Secondly, compared with the two-branch model, the three-branch model will distract attention, resulting in the learned features not being concentrated enough, and cannot be concentrated on discriminating features, resulting in performance degradation. From the experimental results in the table, it can be seen that the effect of using one and four convolution of two branches is the best, and mAP and Rank1 are 1.8% and 1.2% higher than the baseline, respectively.

Table 5. The Multi-Scale Model areas highlighted in dark green in the figure use different scales to explore the optimal setting of the MAV.

Scales	mAP	Rank1
Conv1 & 2	56.4	94.4
Conv1 & 3	56.9	94.7
Conv1 & 4	57.6	95.2
Conv1 & 5	57.1	94.4
Conv2 & 3	54.8	93.2
Conv2 & 4	56.1	93.7
Conv2 & 5	53.8	93.4
Conv1 & 2 & 3	53.9	92.3
Conv1 & 2 & 4	54.2	93.8
Conv1 & 3 & 4	50.7	91.0

4.4.4. The Effectiveness on Different Depths Across View Alignment Model

We analyzed in detail the effect of different iteration depths of AE on model performance in AVAM. The experimental results, as shown in Table 6, reveal a remarkable phenomenon: when AE is iterated three times, the model achieves optimal performance on mAP and Rank1 (R1), which is of 58.2% and 96.9%, respectively. This result shows that the model can effectively capture complex transformations in image features with appropriate AE iteration depth, thus improving the performance of recognition tasks. However, if the number of AE iterations is insufficient, the model may lack sufficient representational power to cope with the diversity and complexity of features, resulting in limited performance. On the contrary, excessive iteration can cause the model to fall into the dilemma of overfitting, making the model too sensitive to the noise and outliers in the training data, and ignoring the importance of generalization ability.

Table 6. Effect of different iteration depth of AE on model performance in AVAM.

Depths	mAP	Rank1
1	55.7	94.6
2	57.6	95.2
3	58.2	96.9
4	55.7	93.2

In addition, we observed that, with the increase in the number of AE iterations, the model performance first improves and then becomes stable, which may eventually damage the generalization of the model due to the increase in model complexity. According to experiments, the optimal number of iterations of AE in AVAM is three, which provides important experience and guidance for designing and optimizing attention enhancement mechanisms in similar tasks in the future. Our research highlights finding the right depth of iteration in model design to achieve the best balance between feature capture capabilities and generalization capabilities.

4.4.5. Visualization of Model Retrieval Results

To underscore the enhanced performance of our model, Figure 7 presents a comparative visualization of the baseline and our approach's retrieval outcomes on the AG1 dataset, focusing on the R1, Rank5 (R5), and R10 metrics. The initial row corresponds to the results yielded by the baseline technique, whereas the subsequent row depicts the results of our model. Correctly identified samples are denoted by a green-bordered image, whereas erroneously retrieved samples are indicated with a red border. Discrepancies between the mistaken samples and the query image are highlighted using a red circular marker.

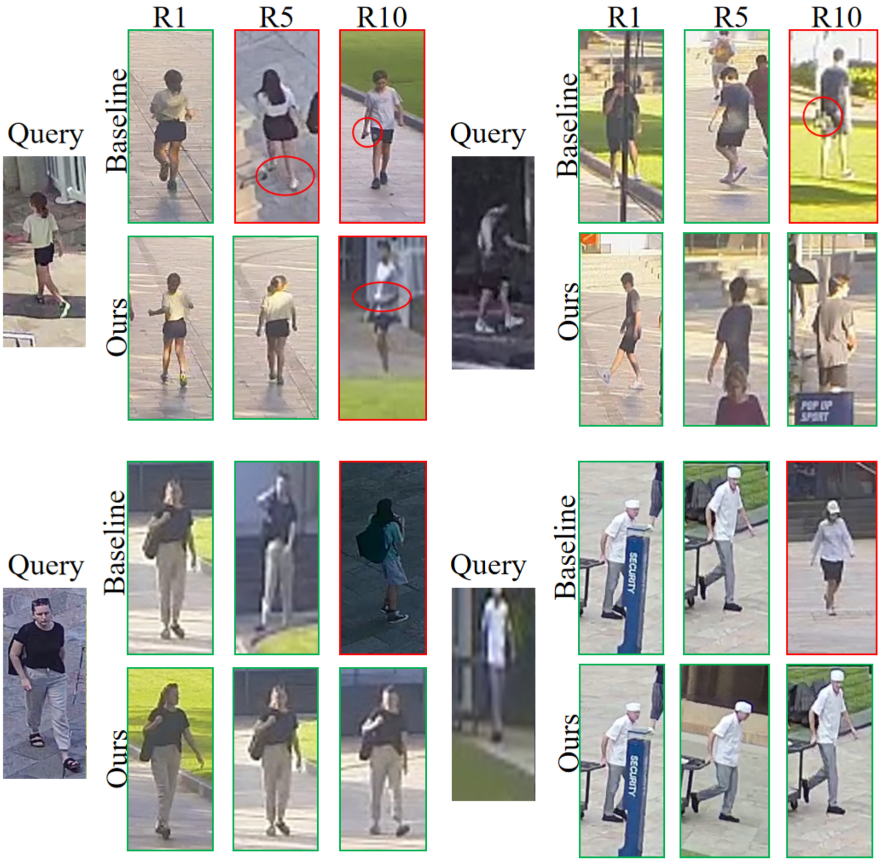

Figure 7. Visualization of baseline and R1, R5, R10 retrieval results on AG1 datasets. The major errors in the picture are circled in red.

As can be observed from Figure 7, the baseline attention is usually focused on the appearance of the character, such as clothing color, body type, etc. As a result, there are some negative matches due to similarities in posture and lighting. For example, the shoes worn by the second character in the first row. The Query image will be misjudged as a white shoe due to the reflection in sunlight, resulting in a negative sample. Our proposed MAVNet increases the diversity of features and makes it easier to distinguish between similar-looking characters. In addition, MAVNet uses a feature alignment module to enhance the alignment of focus on key areas of different perspectives, so that the model has better performance in distinguishing characters. Overall, MAVNet improves the model's ability to capture different detailed features of a person, thereby improving the accuracy and robustness of ground-to-air cross-perspective Re-ID.

5. Conclusions

In this paper, we propose a new ground–air cross-perspective Re-ID task. To address challenges such as scale variation and perspective differences, we first propose a new MAVNet for extracting distinguishing features that are robust to perspective changes. Secondly, in order to make use of multi-scale features reasonably and flexibly, we propose a multi-scale module with cross-view. Finally, we improve the ability of the model to distinguish positive and negative samples by using the alignment module to maintain the feature differentiation in the multi-scale feature fusion. Our proposed MAVNet can achieve SOTA performance.

Author Contributions: Conceptualization, Y.P. ; Methodology, W.Y. and Z.Y.; Visualization, W.L. and B.H.; Software, W.Y. and H.H.; Validation, W.Y and H.H.; Resources, W.Y. and H.H.; Writing—review and editing, H.H. and W.L.; Writing—original draft, W.Y.; Formal analysis, Y.P. and Z.Y.; Project

administration, Y.P. and B.H.; 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 (No. WJ20211A030131); Research and innovation team project of Engineering University of PAP (KYTD202306); Applied research and promotion project of Engineering University of PAP Research (WYY202304); Graduate research funding project (JYWJ2023C010); Graduate research funding project (JYWJ2023B003).

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

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

References

1. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
2. 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.
3. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
4. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 652–662. [CrossRef] [PubMed]
5. Chen, W.; Chen, J.; Zeb, A.; Yang, S.; Zhang, D. Mobile convolution neural network for the recognition of potato leaf disease images. *Multimed. Tools Appl.* **2022**, *81*, 20797–20816. [CrossRef]
6. Tan, M.; Le, Q.V. Mixconv: Mixed depthwise convolutional kernels. *arXiv* **2019**, arXiv:1907.09595.
7. Zhang, X.; Gao, X.; He, L.; Lu, W. MSCAN: Multimodal self-and-collaborative attention network for image aesthetic prediction tasks. *Neurocomputing* **2021**, *430*, 14–23. [CrossRef]
8. Yu, H.X.; Zheng, W.S.; Wu, A.; Guo, X.; Gong, S.; Lai, J.H. Unsupervised person re-identification by soft multilabel learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2148–2157.
9. 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.
10. Kumar, D.; Siva, P.; Marchwica, P.; Wong, A. Fairest of them all: Establishing a strong baseline for cross-domain person reid. *arXiv* **2019**, arXiv:1907.12016.
11. Zhang, S.; Zhang, Q.; Yang, Y.; Wei, X.; Wang, P.; Jiao, B.; Zhang, Y. Person re-identification in aerial imagery. *IEEE Trans. Multimed.* **2020**, *23*, 281–291. [CrossRef]
12. Li, T.; Liu, J.; Zhang, W.; Ni, Y.; Wang, W.; Li, Z. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 16266–16275.
13. Nguyen, K.; Fookes, C.; Sridharan, S.; Liu, F.; Liu, X.; Ross, A.; Michalski, D.; Nguyen, H.; Deb, D.; Kothari, M.; et al. Ag-reid 2023: Aerial-ground person re-identification challenge results. In Proceedings of the 2023 IEEE International Joint Conference on Biometrics (IJCB), Ljubljana, Slovenia, 25–28 September 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–10.
14. Diederik, P.K. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
15. Rao, Y.; Chen, G.; Lu, J.; Zhou, J. Counterfactual attention learning for fine-grained visual categorization and re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 1025–1034.
16. Gu, J.; Wang, K.; Luo, H.; Chen, C.; Jiang, W.; Fang, Y.; Zhang, S.; You, Y.; Zhao, J. Msinet: Twins contrastive search of multi-scale interaction for object reid. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 19243–19253.
17. Luo, H.; Jiang, W.; Gu, Y.; Liu, F.; Liao, X.; Lai, S.; Gu, J. A strong baseline and batch normalization neck for deep person re-identification. *IEEE Trans. Multimed.* **2019**, *22*, 2597–2609. [CrossRef]
18. Luo, C.; Song, C.; Zhang, Z. Generalizing person re-identification by camera-aware invariance learning and cross-domain mixup. In *Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020*; Proceedings, Part XV 16; Springer: Cham, Switzerland, 2020; pp. 224–241.

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.