remote sensing

MDPI

Article

A Faster and More Effective Cross-View Matching Method of UAV and Satellite Images for UAV Geolocalization

Jiedong Zhuang, Ming Dai, Xuruoyan Chen and Enhui Zheng *

Unmanned System Application Technology Research Institute, China Jiliang University, Hangzhou 310018, China; P1901085287@cjlu.edu.cn (J.Z.); s20010802003@cjlu.edu.cn (M.D.); p1901085206@cjlu.edu.cn (X.C.)
* Correspondence: ehzheng@cjlu.edu.cn

Abstract: Cross-view geolocalization matches the same target in different images from various views, such as views of unmanned aerial vehicles (UAVs) and satellites, which is a key technology for UAVs to autonomously locate and navigate without a positioning system (e.g., GPS and GNSS). The most challenging aspect in this area is the shifting of targets and nonuniform scales among different views. Published methods focus on extracting coarse features from parts of images, but neglect the relationship between different views, and the influence of scale and shifting. To bridge this gap, an effective network is proposed with well-designed structures, referred to as multiscale block attention (MSBA), based on a local pattern network. MSBA cuts images into several parts with different scales, among which self-attention is applied to make feature extraction more efficient. The features of different views are extracted by a multibranch structure, which was designed to make different branches learn from each other, leading to a more subtle relationship between views. The method was implemented with the newest UAV-based geolocalization dataset. Compared with the existing state-of-the-art (SOTA) method, MSBA accuracy improved by almost 10% when the inference time was equal to that of the SOTA method; when the accuracy of MSBA was the same as that of the SOTA method, inference time was shortened by 30%.

Keywords: cross-view image matching; geolocalization; UAV image localization; deep neural network

Citation: Zhuang, J.; Dai, M.; Chen, X.; Zheng, E. A Faster and More Effective Cross-View Matching Method of UAV and Satellite Images for UAV Geolocalization. *Remote Sens.* **2021**, *13*, 3979. https://doi.org/10.3390/rs13193979

Academic Editor: Eufemia Tarantino

Received: 21 August 2021
Accepted: 1 October 2021
Published: 5 October 2021

1. Introduction

Unmanned aerial vehicles (UAVs) have rapidly developed in recent years and gradually become the main platform for remote-sensing image acquisition due to their convenient operation and strong data-collection capabilities. Alexander et al. [1] used data from a UAV to locate trees in a rainforest. Ammour et al. [2] proposed a deep-learning method of detecting cars in UAV imagery. Deng et al. [3] applied UAV-based multispectral remote sensing to precision agriculture. UAV application involves many fields, such as photogrammetry, agriculture, and mapping [4–7]. However, the current positioning and navigation of UAVs mainly rely on positioning systems such as GPS and GNSS. How to achieve autonomous positioning and navigation of UAVs without the assistance of positioning systems is a challenging task. Cross-view image-matching technology matches a satellite image with geographic location tags and UAV images without geographic location tags, so as to realize UAV image positioning and navigation.

Cross-view geolocalization has been a research focus in recent years due to the huge application value of geolocalization. Many previous approaches focus on extracting handcrafted features from images [8–11]. With the rapid development of the performance of convolutional neural networks, many studies used a CNN to automatically extract robust features in images. To bridge the gap between images of different views, the Siamese network was proposed [12], which is helpful for a model to learn viewpoint-invariant features. Liu et al. [13] designed a network to encode pixels in images by orientation. Hu et al. [14] proposed a Siamese-like architecture, acquiring more robust feature description

by using NetVLAD [15]. Another line of works focused on metric learning and studied various loss functions to learn discriminative features. In order to close the gap between the learned and true distribution, Liu et al. [16] proposed stochastic attraction and repulsion embedding loss. Vo et al. [17] showed that explicit orientation supervision can improve location prediction accuracy by orientation regression loss. Some other works focused on domain alignment. Shi et al. [18] applied polar transform to warp aerial images, and realize the alignment between aerial and ground views. Further, they designed a DSM method [19] by adopting a dynamic similarity-matching network to estimate cross-view orientation alignment during localization. Most of the above work was carried out on two datasets, CVUSA [20] and CVACT [13], which are used for matching ground-view and satellite-view images. Zheng et al. proposed UAV-based dataset University-1652 [21], and examined image-retrieval tasks from a classification perspective. The model optimized by instance loss [22,23] could learn more discriminative embeddings than ranking loss could. In a recent work, inspired by partition strategies [24–27], Wang et al. [28] proposed a local pattern network (LPN) concentrating on matching drone- and satellite-view images based on University-1652. LPN incorporated contextual information in cross-view geolocalization, and applied a rotation-invariant square-ring feature partition strategy to enable the network to focus on auxiliary information such as houses, roads, and trees, locating around the geographic target, which was a breakthrough in the field of cross-view geolocalization.

The goal of attention is to focus on salient features and suppress irrelevant features [29]. Research in the field of Re-id is quite extensive [30–34], but rare in cross-view geolocalization. One line of works used convolutional operations with small receptive fields on feature maps. Inspired by SE-Net [35], Wang et al. [36] proposed a full attention block (FAB) to prevent the loss of spatial information. To maximize complementary information, Li et al. [37] formulated the HA-CNN for the joint learning of soft pixel attention and hard regional attention. Chen et al. [38] considered previous works, focused on coarse attention, and explicitly enhanced the discrimination and richness of attention knowledge. Another line of works widened the receptive field to introduce more contextual information. Wang et al. [39] stacked attention modules to generate attention-aware features, and applied attention residual learning to train deep networks. Yang et al. [40] designed an intra-attention network search for informative and discriminative regions in images. Zhang et al. [41] inserted a nonlocal block [42] before the encoder–decoder style attention module to enable attention learning based on globally refined features. In a recent work, Zhang et al. [29] mined the relation representation between feature nodes to learn semantics and infer attention.

Since Hinton et al. [43] proposed the distillation of knowledge in 2015, joint learning has been rapidly developing in the field of deep learning. Borrowing this idea, Romero et al. [44] used large, powerful, and easy-to-train networks to guide small but harder-to-train networks. JIM et al. [45] regarded the flow between layers computed by the inner product of feature maps between layers as distilled knowledge. In a recent work, Wang et al. [46] encoded the features into a pyramid structure, and encouraged shorter codes to mimic longer codes by self-distillation. Different from the above distillation-based method, Zhang et al. [47] presented a deep mutual learning (DML) strategy with an ensemble of students learning collaboratively and teaching each other throughout the training process.

However, most of above cross-view methods focus on global information, and few focus on contextual information. Existing hard part-based representation learning strategies ignore the offset and scale of the location, and there are few attention mechanisms specifically designed for cross-view geolocalization. To overcome the above shortcomings, we achieved the following improvements:

1. The existing partition strategy was improved, such that the proposed model is more robust to offset and scale changes. Different from existing works, the model was divided into global and local branches, which were trained together to combine the whole and parts (see Sections 2.2–2.4).

2. An attention mechanism is proposed for cross-view geolocalization and the partition strategy. Specifically, the attention module finds the relationship between regions and makes each region pay attention to different features (see Section 2.5).

3. We minimized KL divergence loss to narrow difference between domains (see Section 2.6).

4. As a result, the model achieved outstanding performance. The test results on the benchmark dataset showed that the accuracy of the model in the task of matching drone and satellite images greatly exceeded that of the best existing model. Moreover, compared with the SOTA method, the model reduced inference time by 30% and achieved the same level of accuracy. When inference time was almost the same, the accuracy of the model was much ahead of that of the SOTA method (see Section 3).

2. Materials and Methods

2.1. Datasets and Evaluation Indicators

The research was conducted on University-1652, released by Zheng [21]. University-1652 contains 1652 locations or so-called geographic targets from 72 universities from all over the world. Each geographic target contains three views: satellite, drone, and street views. Each target has 1 satellite-view image, over 50 drone-view images from different filming angles and heights, and a few street-view images. Satellite and drone views were selected for this research, and the performance of the method was tested in matching these two views. Method performance is mainly reflected in two tasks, Drone → Satellite and Satellite → Drone. Specifically, the purpose of the former is giving a drone image and finding the drone image of the same place; the purpose of the latter is giving a satellite image and finding the K satellite images of the corresponding place. The details of the query and gallery in the datasets are shown in Table 1. In the testing dataset of the Drone → Satellite task, there was only one true-matched satellite-view image for each drone-view image.

Table 1. Statistics experimental data. Including the image number of query set and gallery set for different geolocalization tasks.

Training Dataset		
Views	**Numbers of Buildings**	**Numbers of Images**
Drone	701	37,854
Satellite	701	701
Testing Dataset		
Views	**Numbers of Buildings**	**Numbers of Images**
$\text{Drone}_{\text{query}}$	701	37,854
$\text{Satellite}_{\text{query}}$	701	701
$\text{Drone}_{\text{gallery}}$	951	51,355
$\text{Satellite}_{\text{gallery}}$	951	951

Satellite-view images with geotags were from Google Maps. Google Maps images have a similar scale to that of drone-view images and high spatial resolutions (from level 18 to 20, the spatial resolution ranges from 1.07 to 0.27 m).

Due to airspace control and high cost, it is very difficult to collect a large number of real drone-view images, so the drone-view images were simulated by the 3D model provided by Google Earth. The view in the 3D model spirally descends, and the height of view from 256 to 121.5 m, while images were recorded at regular intervals, so as to obtain a large number of drone images close to the real world.

In the field of cross-view matching and image retrieval, Recall@K (R@K) and average precision (AP) are essential evaluation indicators. R@K refers to the ratio of the number of matched images in the top-K ranking list, and measures the recall rate of the retrieval

system. AP measures the precision of the retrieval system. These two are used to evaluate the performance of methods.

2.2. Backbone Structure

Multiscale block attention network (MSBA) consists of two major branches, namely, satellite and drone views, and the street-view branch was removed, which is different from other mainstream methods [21,28,48]. Focusing on both performance and efficiency, Resnet50 [49] is used as the backbone to extract image features. Alternative backbones with other classical networks such as VGG [50] or new networks with distinguished performance are also candidates to achieve better results. The average pooling layer and classification layer behind layer4 in Resnet50 were both removed. In order to compare with other methods, schematically taking an image with the same size of 256 × 256 as that of the input of the network, the entire network structure and the forward propagation process are shown in Figure 1. The whole structure was divided into two branches, the drone- and satellite-view branches, and they share the weights of the backbone. Each branch contains global (branch marked with a red arrow) and local (branch marked with a yellow arrow) branches. In the global branch, feature maps after layers 3 and 4 (details in Section 2.4) are concatenated, and the next is sent to the max pooling layer. In the local branch, the feature map after layer 4 is divided into six blocks (details in Section 2.3), and average pooling and attention operations are then performed (details in Section 2.5) on each block. All pooled features are sent to the classifier module, including the fully connected, batch normalization, dropout, and classification layers. In the training stage, the model is optimized by minimizing cross-entropy loss and KL divergence loss (details in Section 2.6). During testing, the classification layer is removed from the classifier module, and all features are concatenated to calculate the Euclidean distance of the image in the feature space.

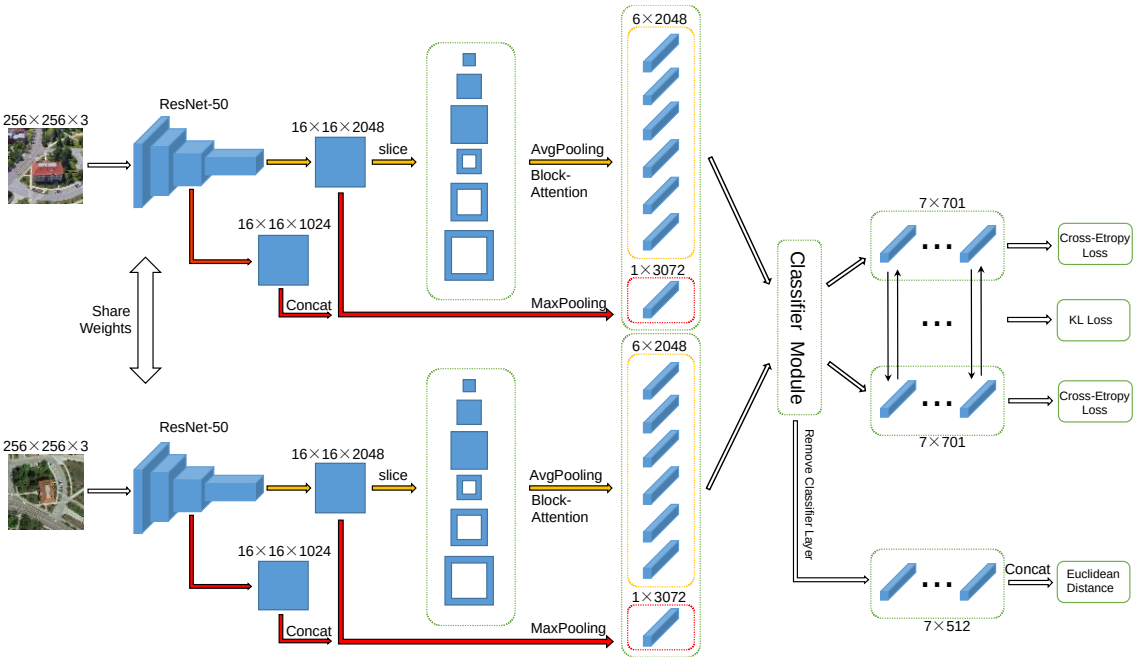


Figure 1. Framework of proposed MSBA.

2.3. Multiscale Block

Considering the importance of contextual information, the square-ring partition strategy is applied to divide feature maps, just like that of LPN. The square-ring feature partition strategy can fully mine contextual information in an image, but is also accompanied with the frequent problems of offset and scale uncertainty of geographic targets. As shown in

Figure 2, the scale of the geographic target in satellite query and drone gallery images is obviously different, and the geographic target is offset from the center of the drone image. The square-ring feature partition strategy is taken for these two images, and the three parts close to the center of the image are used to illustrate the problem of offset and scale uncertainty (Figure 2A). The three parts in the two images were obviously misaligned, and a large error was caused by using the distance calculation method in Figure 2A to measure the similarity between the two parts. In this case, the method in Figure 2B may be more reasonable. In summary, a hard partition strategy is flawed in dealing with changes in scale and offset, and is unable to balance local and global information, which is equally important. Therefore, we address this problem by improving the square-ring feature partition strategy and adding two global blocks with different scales on its basis, which is helpful in resisting offset and scale changes. The six segmented feature maps (see Figure 3) are denoted as $f_i^j (i \in [1,6]; j \in [1,2])$. The pooling operation can be formulated as

$$z_i^j = P_{avg}\left(f_i^j\right) \tag{1}$$

where P_{avg} stands for the average pooling operation, and z_i^j stands for the result after the pooling operation.

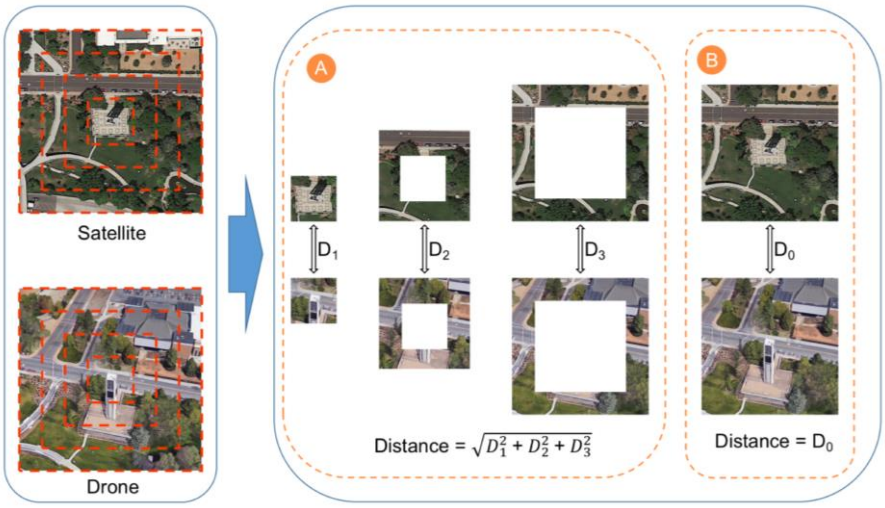


Figure 2. Square-ring feature partition strategy. (**A**,**B**) Two ways to calculate distance between image features.

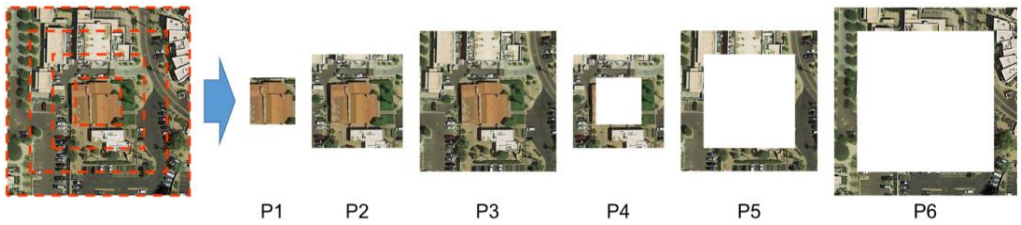


Figure 3. Improved block strategy. Feature map is divided into six blocks of the above shape.

2.4. Global Branch

In order to extract more robust features from the image, a global branch is added after the backbone. Figure 4 shows that, if only local information is considered to train the network, the attention on the central geographic target may be neglected.

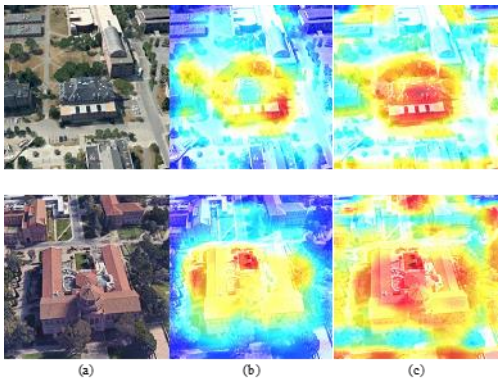


Figure 4. Grad-CAM [51] visualization of heatmaps: (**a**) Original images; (**b**) LPN methods; (**c**) MSBA methods.

Shallow layers in neural networks tend to extract information such as structural edges. As the network deepens, extracted information tends to be semantic. In cross-view tasks, such as geographic target matching, the shape, structure, and edges of buildings that concern the shallow layers are also important guidance information. However, most existing methods employ deep features extracted by the last layer of backbones. Therefore, in the global branch, the shallow and deep features of the backbone network are extracted together. The difference from the local branch is that the max pooling operation is applied after layers 3 and 4. The operation can be formulated as

$$w^j = P_{max}\left(g^j\right) \tag{2}$$

where g^j stands for the global feature map, and P_{max} stands for the max pooling operation.

2.5. Block Attention (BA)

In drone and satellite images, geographic targets are often close to the center, while trees and roads are distributed outside. The farther an object leaves from the center, the smaller the correlation it has with geographic targets. On this basis, an attention mechanism was designed to reinforce the salient features in different regions and suppress some redundant features.

When feature extraction is simultaneously performed on multiple local feature blocks, extracted features are redundant in each dimension. For the purpose of deredundancy, and letting the local feature blocks perform their duties, a block self-attention module was added after the feature map, which is partitioned by the local branch. The entire process of BA is shown in Figure 5. BA is formulated as follows:

$$Z^j = \mathcal{C}\left(z_i^j\right)\ i \in\ [1,6] \tag{3}$$

$$\mathcal{X}^j = Softmax\left(Z^j\right) \odot Z^j \tag{4}$$

where $\mathcal{C}$ stands for the concatenate operation, $Softmax$ stands for performing a softmax operation in the N direction, and $\odot$ stands for multiply corresponding elements between matrices.

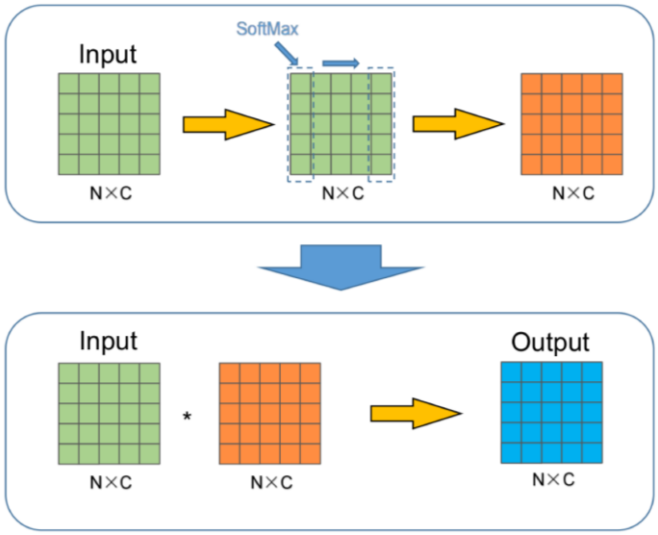


Figure 5. Process of block attention. In MSBA method, N = 6 and C = 2048.

2.6. Loss Function and Improvement in Inference

Each part of the features is sent to the fully connected layer, and softmax processing is performed on the output of that layer to normalize the result to a feature space with the value range from 0 to 1. The optimization goal is that, in this feature space, feature vectors of the same geographic target are closer, while the feature vectors of different geographic targets are farther away. The network is optimized by the cross-entropy loss function and it can be formulated as:

$$CE(p, y) = \begin{cases} -\log(p) \; y = 1 \\ -\log(1-p) \; otherwise \end{cases} \tag{5}$$

where p stands for forecast result, and y stands for the ground-truth label.

Model self-distillation and mutual learning are familiar neural-network training techniques [43,47], but they are rarely used in the cross-view geolocalization with domain gaps between drone images and satellite images. Specifically, the twin networks learn from each other, and study the implicit relationship between satellite- and drone-view images. The corresponding local and global features in the two branches are subjected to a softmax operation; KL divergence is then calculated and used as loss to train the network. The KL distance from p_1 to p_2 is computed as:

$$D_{KL}(p_2||p_1) = \sum_{m=1}^{M} p_2^m log \frac{p_2^m}{p_1^m} \tag{6}$$

where p_1, p_2 stand for forecast result of the two branches, respectively.

According to the characteristics of the satellite- and drone-view images, geographic targets are generally located close to the center of the image. Therefore, when calculating the distance between features in the inference stage, more attention should be paid to the center of the image. In the inference stage, the feature vector is multiplied corresponding to each feature block by a weight, and then concatenated into a total feature vector to calculate the distance between images. The reweighting process can be formulated as:

$$F = \mathcal{C}(\beta_i * d_i) \; i \in [1, 7] \tag{7}$$

where β_i stands for weights added on the feature vector d_i, and $\mathcal{C}$ stands for the concat operation.

2.7. Implementation

Unless otherwise specified, all training and testing images were resized to 256 × 256. Training images were augmented with random flipping and random cropping. Batch size was set to 8 due to the limitation of memory. The backbone was initialized by loading the weights pretrained on ImageNet by ResNet50. The stride of the final downsampling layer was adjusted from 2 to 1 to increase the size of the feature map output by the backbone, which is a common trick in image retrieval. SGD was chosen as the optimizer with momentum of 0.9 and weight decay of 5×10^{-4} to train the model. The model was trained for 140 epochs; the initial learning rate was 1×10^{-4} for the backbone layers, and 1×10^{-3} for the other layers. After 80 epochs, the learning rate dropped to one-tenth of the original. In the inference stage, the similarity between images was evaluated by calculating the Euclidean distance between image feature vectors. All experiments were performed with an Nvidia 2070s GPU using the PyTorch deep-learning framework with FP16 training.

3. Experiments Results

3.1. Comparison with the State of the Art

In Table 2, MSBA method is compared with other methods on University-1652. The method achieved 82.33% R@1 accuracy and 84.78% AP on the task of Drone → Satellite, 90.58% R@1 accuracy and 81.61% AP on the task of Satellite → Drone with standard input (image size of 256 × 256). When an image with 384 × 384 size was used as input, the method achieved 86.61% R@1 accuracy and 88.55% AP, and 92.15% R@1 accuracy and 84.45% AP on two tasks. The performance of the method greatly surpassed that of existing competitive models, which was nearly 10% higher than that of the existing best-performing method, i.e., LPN in some indicators.

Table 2. Comparison of state-of-the-art results reported on University-1652. s384, input size.

Method	Publication	Backbone	Drone → Satellite		Satellite → Drone	
			R@1	AP	R@1	AP
Soft margin triplet loss [13]	CVPR'19	VGG16	53.21	58.03	65.62	54.47
Instance loss [22]	TOMM'20	ResNet-50	58.23	62.91	74.47	59.45
Instance loss + verification loss [16,23]	TOMM'17	ResNet-50	61.30	65.68	75.04	62.87
Instance loss + gem pooling [52]	TPAMI'18	ResNet-50	65.32	69.61	79.03	65.35
LCM [48]	Remote Sens'20	ResNet-50	66.65	70.82	79.89	65.38
LPN [28]	TCSVT'21	ResNet-50	74.16	77.39	85.16	73.68
Ours	-	ResNet-50	82.33	84.78	90.58	81.61
Ours(s384)	-	ResNet-50	86.61	88.55	92.15	84.45

3.2. Ablation Study of Methods

As shown in Table 3, ablation experiments were performed on the proposed method. This provided +6.16% R@1 and +5.58%AP on the task of Drone → Satellite, and +4.56%R@1 and +7.24%AP on Satellite → Drone. In the experiment, branch mutual learning and block attention are two significant tricks to improve performance, contributing approximately 3% points of improvement in both tasks. In the inference stage, assigning different weights to all the feature vectors in the local and global branches, and then calculating the Euclidean distance between images can also improve the retrieval performance of the model, especially when a larger weight is assigned to the feature vector of the global branch, which is demonstrated in Figure 6 by a performance test of R@1 on a weight from 1 to 2. As shown in Figure 6, the performance of R@1 temporarily improved with a weight. When weight = 1.5, the model had the best performance. When the weights continued to increase, the performance of the model began to worsen. So we chose weight = 1.5 as the parameters of the model during inference for the global branch feature vector.

Table 3. Results of ablation experiments of proposed method.

Method	Drone → Satellite R@1 AP	Satellite → Drone R@1 AP
LPN	74.16 77.39	85.16 73.68
Baseline(ours)	76.17 79.20	86.02 74.37
+Multiscale	77.18 79.94	85.93 75.12
+Global branch	78.47 81.42	86.16 77.17
+Block attention	79.55 82.30	87.87 79.94
+KL loss	81.29 83.85	89.73 80.64
+Reweights	82.33 84.78	90.58 81.61

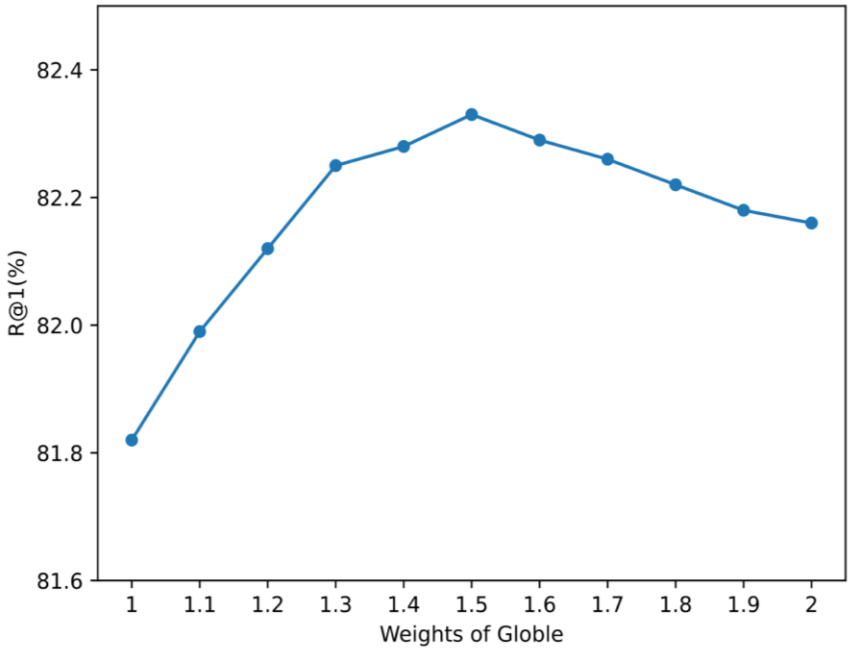


Figure 6. Effect of global branch weights.

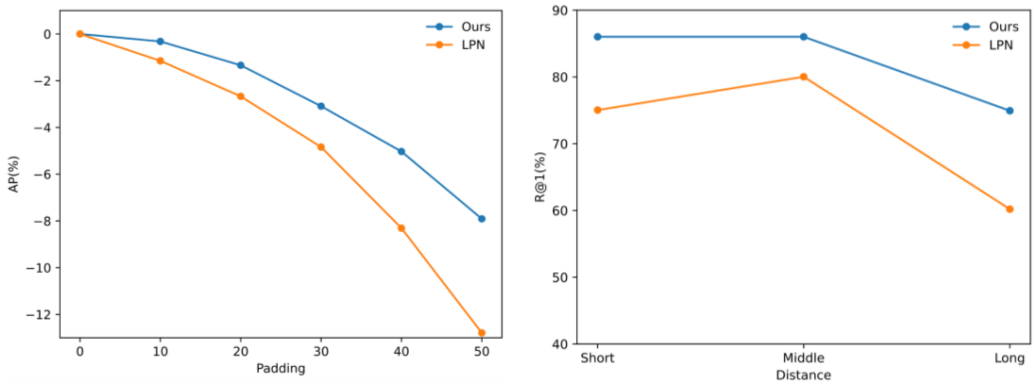
3.3. Offset and Scale Ablation

In order to confirm whether the robustness of the model to offset and scale changes was really improved compared to the current mainstream methods, a set of ablation experiments were designed. First, the antioffset of the model was tested, and the result is displayed in Table 4. The image was shifted from 0 to 50 pixels to the right to offset the geographic target from the center. In order to ensure the reliability of the results, an image with a size of 384 × 384 was used as the input of the model. Experimental results showed that, when the offset slowly increased from 0, the performance of the model did not significantly change. Even when the offset reached 50, the accuracy of the model on the two tasks was still very competitive. Figure 7a shows the comparison result of the model with the current model with good performance; with the increase in offset, the attenuation of the model was much smaller than that of the existing model, which also proves that the model was more robust to the offset. Second, we tested the scale change on the task of Drone → Satellite. Drone-query images were split into short, medium, and long groups on the basis of the distance between drone and geographic target. The model performed excellently with short and medium distances. When the distance was long, the performance of the model was somewhat degraded, but it was still considered to be a good result (Table 5). The comparison of the model proposed with the existing model is shown in Figure 7b. The performance of the existing model fluctuated with the distance between drone and geographic target changing from short to medium, and the distance sequentially increased, while its performance significantly decreased. Overall, the model showed better

adaptability to different interferences and maintained excellent performance with higher accuracy than the existing SOTA method did.

Table 4. Results of ablation experiments of shifting query images during inference.

Padding Pixel	Drone → Satellite R@1 AP	Satellite → Drone R@1 AP
0	86.61 88.55	92.15 84.45
10	86.22 88.23	92.01 84.24
20	85.02 87.21	90.87 83.43
30	83.95 85.46	89.44 81.95
40	81.48 83.52	88.30 79.64
50	77.53 80.64	85.59 75.84



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

Figure 7. (**a**) Effect of shifting in query images. (**b**) Effect of distance to geographic target in query images.

Table 5. Results of ablation experiments of using drone images with different distance to geographic target to conduct retrieval. "All" stands for using all drone-views query images.

Distance	Drone → Satellite R@1 AP
All	82.33 84.78
Short	86.02 87.93
Middle	86.02 88.09
Long	74.94 78.27

3.4. Input Size Ablation

Low-resolution images negatively impact the performance of the model, but some real-world applications with computing resource limitations or requirements for operating efficiency require to use low-resolution images and lose accuracy for matching drone images with satellite images. A set of ablation experiments were used to observe changes in model performance when images of different resolutions were used as input. As shown in Table 6, when the input image resolution changed from 224 to 512, model performance improved. Only in the task of Satellite → Drone, when the resolution changed from 384 to 512, did performance decrease. This shows that the proposed model could adapt to resolution input, that is, it could be applied to a wide range of real-world scenarios. The performance of the proposed model was compared with that of the existing model at different resolutions, as shown in Figure 8. LPN performance very obviously degraded when the resolution was reduced to 224. Generally speaking, the model had good competitiveness in application scenarios with time and space complexity constraints.

Table 6. Results of ablation experiments of different input sizes of images during training and testing.

Input Image Size	Drone → Satellite R@1 AP	Satellite → Drone R@1 AP
224	78.09 80.95	87.30 77.37
256	82.33 84.78	90.58 81.61
320	85.40 87.47	92.01 84.23
384	86.61 88.55	92.15 84.45
512	86.69 88.66	92.01 84.45

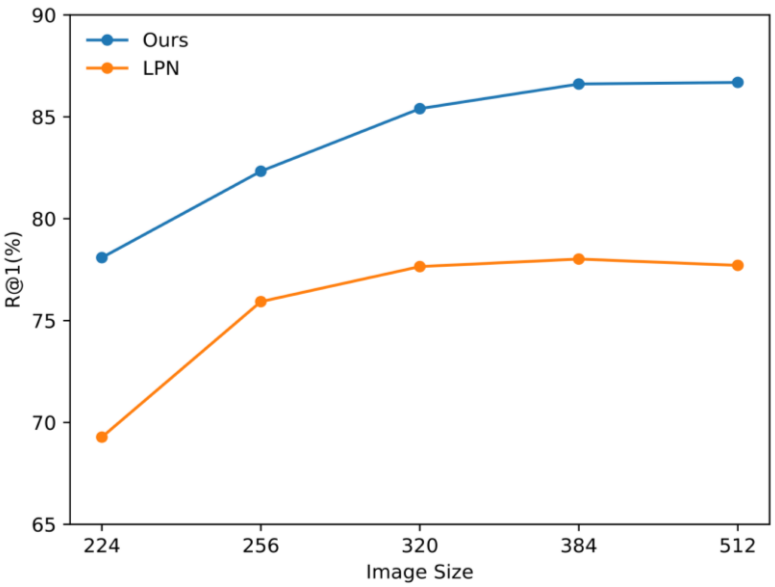


Figure 8. Effect of input size of images.

3.5. Inference with Different Branches

In the inference stage, feature maps from the global and the local branches are split into seven blocks. As shown in Table 7, their combinations were selected to calculate the Euclidean distance between image features. Results showed that, when each feature block was individually tested, the model did not perform well on the two tasks, but when they were combined, the model reached a fairly high level of performance, which confirmed the contextual information and the importance of multibranch training. When only the global branch was used for inference, the model reached R@1 73.59%, AP 76.93% on the Drone → Satellite task, and R@1 83.74%, AP 72.12% on the Satellite → Drone task, which almost achieved the best performance of the existing models. This shows that the global and local branches can be trained together to promote each other's capability to extract robust features. The last column of the Table 7 shows the model inference time compared with that of LPN. The proposed model can greatly reduce inference time while ensuring accuracy, and accuracy is greatly improved when inference time is almost the same. Moreover, if only the global branch is used in the inference phase, much time and many computing resources can be saved, which makes the model very suitable for real application scenarios.

Table 7. Results of ablation experiments of using different parts during inference. "P1–P6" means using only one part in Figure 4. "Global" means only using global branch. "All" means the combination of all of them. "RW" means reweights.

Part Combination	Drone → Satellite R@1 AP	Satellite → Drone R@1 AP	Inference Time
P1	51.73 56.64	70.76 51.31	—
P2	59.66 63.77	77.75 58.85	—
P3	52.89 57.18	78.32 53.17	—
P4	63.49 67.70	79.32 62.95	—
P5	63.30 67.37	78.17 62.53	—
P6	55.09 55.09	78.03 55.28	—
Global	73.59 76.93	84.74 73.12	0.72×
All	81.29 83.85	89.73 80.64	1.05×
All(RW)	82.33 84.78	90.58 81.61	1.05×
LPN	74.16 77.39	85.16 73.68	1.00×

3.6. Result Visualization

To further prove the reliability of proposed method, visualized results are displayed in Figure 9, showing the heatmaps generated by LPN and MSBA on some testing images. MSBA focuses on more regions in the image than those of LPN, especially the location of geographic targets. This phenomenon may be because of adding global branches and giving different weights to different regions in the training, but LPN treats different regions equally. The retrieval results of Satellite → Drone and Drone → Satellite tasks are also visualized. Figure 10 shows that MSBA had a high top-5 hit rate on the task of Satellite → Drone; on the other task, the method could correctly retrieve even in some very similar images.

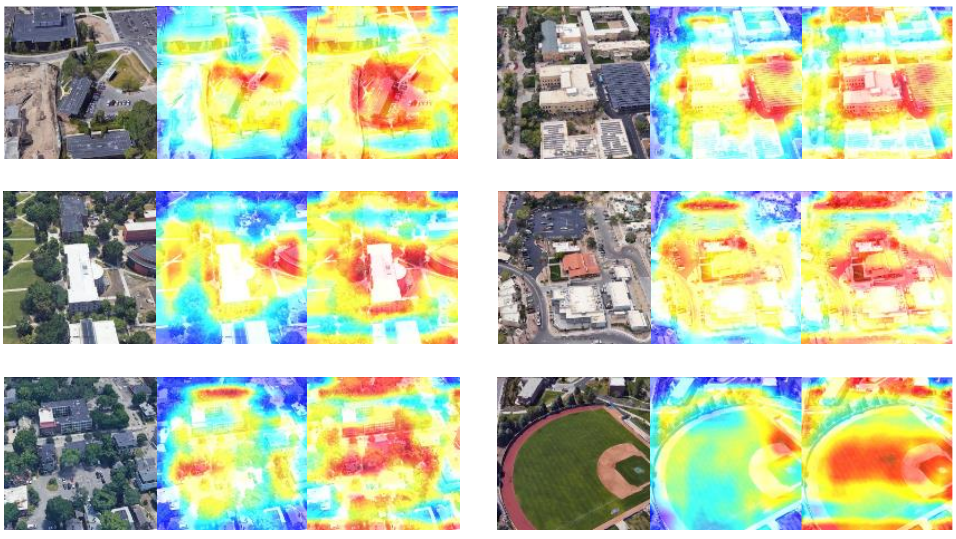


Figure 9. Grad-CAM visualization of attention maps. (**a**) Input images, (**b**) LPN, (**c**) MSBA.

Figure 10. Visualized images retrieval results. (**a**) Top-5 retrieval results of Satellite → Drone. (**b**) Top-5 retrieval results of Drone → Satellite. True matches are in green boxes; false matches are displayed in red boxes.

4. Discussions

Through the experiments results on two tasks (Satellite → Drone and Drone → Satellite), we deeply explored the model's retrieval performance and compared it with that of existing models. The participation of global and local branches in training can make the model pay more attention to important information in the images, such as central geographic landmarks. From the experiments of combining different blocks, the joint training of the global and local branches greatly improved the accuracy of the global branch in inference. In the case of improving the accuracy of SOTA method, inference time was shortened by using global branch alone because the dimension of the global feature was only 512, and the parameter number of the model was reduced. Nevertheless, the experiment of assigning different weights to the features of the global branch in the inference stage shows that, if the features of the global branches are given too large weight, the accuracy of the model is worsened, which reveals that extracted features from each part of the image are essential. In the offset and scale experiments, MSBA was more robust than the existing method. However, when the offset and scale changes were obvious, the performance of the model worsened, so only using multilevel blocks to solve these problems

has limitations. How to solve the problems of offset and scale change is meaningful and promising work, and we will conduct further research on these two tasks.

5. Conclusions

In this paper, we began with the autonomous positioning and navigation of drones without positioning system, leading to two tasks (Satellite → Drone and Drone → Satellite) in cross-view geolocalization. Next, we indicated the shortcomings of the existing method, and proposed a method with better accuracy, which is more robust to offset and scale changes.

Experiment conclusions are as follows: (1) The multiscale division of feature maps can reduce the sensitivity of the model to offset and scale. (2) The joint training of the global and local branches leads the network to more fully mine features in an image. (3) The block attention mechanism and KL divergence loss can further improve the retrieval accuracy of the model. Moreover, the method (MSBA) achieved excellent accuracy and inference speed on drone-based geolocalization benchmarks (University-1652). Due to the widespread existence of offset and scale, and the limitation of computing time in real scenes, MSBA is very competitive in practical applications.

Author Contributions: Conceptualization, J.Z. and E.Z.; methodology, J.Z.; software, M.D.; validation, J.Z., M.D. and X.C.; resources, M.D.; data curation, M.D.; writing—original draft preparation, J.Z.; writing—review and editing, J.Z. and M.D.; visualization, J.Z. and X.C.; supervision, E.Z.; project administration, E.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data sharing not applicable.

Acknowledgments: The authors would like to thank Tianyu Wang from the Intelligent Information Processing Lab, Hangzhou Dianzi University; and Zheng Zhedong from the Reler laboratory; University of Technology, Sydney for providing the Universty-1652 dataset, open-source code, and their contribution in this field. We also thank Jingjing Xiong from China Jiliang University and Jiliang Luo from Huaqiao University for their advice and help for the article.

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

References

1. Alexander, C.; Korstjens, A.H.; Hankinson, E.; Usher, G.; Harrison, N.; Nowak, M.G.; Abdullah, A.; Wich, S.A.; Hill, R.A. locating emergent trees in a tropical rainforest using data from an Unmanned Aerial Vehicle (UAV). *Int. J. Appl. Earth Obs. Geoinf.* **2018**, *72*, 86–90. [CrossRef]
2. Ammour, N.; Alhichri, H.; Bazi, Y.; Benjdira, B.; Alajlan, N.; Zuair, M. Deep learning approach for car detection in UAV imagery. *Remote Sens.* **2017**, *9*, 312. [CrossRef]
3. Deng, L.; Mao, Z.; Li, X.; Hu, Z.; Duan, F.; Yan, Y. UAV-based multispectral remote sensing for precision agriculture: A comparison between different cameras. *ISPRS J. Photogramm. Remote Sens.* **2018**, *146*, 124–136. [CrossRef]
4. Lin, Y.-C.; Cheng, Y.-T.; Zhou, T.; Ravi, R.; Hasheminasab, S.M.; Flatt, J.E.; Troy, C.; Habib, A. Evaluation of UAV LiDAR for mapping coastal environments. *Remote Sens.* **2019**, *11*, 2893. [CrossRef]
5. Yan, Y.; Deng, L.; Liu, X.; Zhu, L. Application of UAV-based multi-angle hyperspectral remote sensing in fine vegetation classification. *Remote Sens.* **2019**, *11*, 2753. [CrossRef]
6. Liu, W.; Yang, M.; Xie, M.; Guo, Z.; Li, E.; Zhang, L.; Pei, T.; Wang, D. Accurate building extraction from fused DSM and UAV images using a chain fully convolutional neural network. *Remote Sens.* **2019**, *11*, 2912. [CrossRef]
7. Ferrer-González, E.; Agüera-Vega, F.; Carvajal-Ramírez, F.; Martínez-Carricondo, P. UAV Photogrammetry accuracy assessment for corridor mapping based on the number and distribution of ground control points. *Remote Sens.* **2020**, *12*, 2447. [CrossRef]
8. Castaldo, F.; Zamir, A.; Angst, R.; Palmieri, F.; Savarese, S. Semantic cross-view matching. In Proceedings of the Workshops of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1044–1052.
9. Lin, T.; Belongie, S.; Hays, J. Cross-view image geolocalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 891–898.
10. Senlet, T.; Elgammal, A. A framework for global vehicle localization using stereo images and satellite and road maps. In Proceedings of the Workshops of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2034–2041.

11. Bansal, M.; Sawhney, H.; Cheng, H.; Daniilidis, K. Geo-localization of street views with aerial image databases. In Proceedings of the ACM International Conference on Multimedia, New York, NY, USA, 28 November 2011; pp. 1125–1128.
12. Chopra, S.; Hadsell, R.; LeCun, Y. Learning a similarity metric discriminatively, with application to face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; pp. 539–546.
13. Liu, L.; Li, H. Lending orientation to neural networks for cross-view geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5624–5633.
14. Hu, S.; Feng, M.; Nguyen, R.M.; Hee Lee, G. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7258–7267.
15. Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; Sivic, J. Netvlad: Cnn architecture for weakly supervised place recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5297–5307.
16. Liu, L.; Li, H.; Dai, Y. Stochastic attraction-repulsion embedding for large scale image localization. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 2570–2579.
17. Vo, N.N.; Hays, J. Localizing and orienting street views using overhead imagery. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 494–509.
18. Shi, Y.; Liu, L.; Yu, X.; Li, H. Spatial-aware feature aggregation for image based cross-view geo-localization. In Proceedings of the Neural Information Processing Systems, Vancouver, VBC, Canada, 8–14 December 2019.
19. Shi, Y.; Yu, X.; Campbell, D.; Li, H. Where am I looking at? Joint location and orientation estimation by cross-view matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 4063–4071.
20. Zhai, M.; Bessinger, Z.; Workman, S.; Jacobs, N. Predicting ground-level scene layout from aerial imagery. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 867–875.
21. Zheng, Z.; Wei, Y.; Yang, Y. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In Proceedings of the ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 1395–1403.
22. Zheng, Z.; Zheng, L.; Garrett, M.; Yang, Y.; Xu, M.; Shen, Y.-D. Dual-path convolutional image-text embeddings with instance loss. *ACM Trans. Multimed. Comput. Commun. Appl.* **2020**, *16*, 1–23. [CrossRef]
23. Zheng, Z.; Zheng, L.; Yang, Y. A Discriminatively Learned CNN Embedding for Person Reidentification. *ACM Trans. Multimedia Comput. Commun. Appl.* **2017**, *14*, 1–20. [CrossRef]
24. Sun, Y.; Zheng, L.; Yang, Y.; Tian, Q.; Wang, S. Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline). In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 501–518.
25. Zheng, L.; Huang, Y.; Lu, H.; Yang, Y. Pose-invariant embedding for deep person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 4500–4509. [CrossRef] [PubMed]
26. Wei, L.; Zhang, S.; Yao, H.; Gao, W.; Tian, Q. Glad: Global– local-alignment descriptor for scalable person re-identification. *IEEE Trans. Multimed.* **2018**, *21*, 986–999. [CrossRef]
27. Zheng, Z.; Zheng, L.; Yang, Y. Pedestrian alignment network for large-scale person re-identification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *29*, 3037–3045. [CrossRef]
28. Wang, T.; Zheng, Z.; Yan, C.; Zhang, J.; Sun, Y.; Zheng, B.; Yang, Y. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Trans. Circuits Syst. Video Technol.* **2021**. [CrossRef]
29. Zhang, Z.; Lan, C.; Zeng, W.; Jin, X.; Chen, Z. Relation-aware global attention for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3183–3192.
30. Zhao, L.; Li, X.; Zhuang, Y.; Wang, J. Deeply-learned part-aligned representations for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3239–3248.
31. Song, C.; Huang, Y.; Ouyang, W.; Wang, L. Mask-guided contrastive attention model for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1179–1188.
32. Suh, Y.; Wang, J.; Tang, S.; Mei, T.; Lee, K.M. Part-aligned bilinear representations for person re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 418–437.
33. Xu, J.; Zhao, R.; Zhu, F.; Wang, H.; Ouyang, W. Attention-aware compositional network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2119–2128.
34. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 907–915.
35. 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.
36. Wang, C.; Zhang, Q.; Huang, C.; Liu, W.; Wang, X. Mancs: A multi-task attentional network with curriculum sampling for person re-identification. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 384–400.

37. Li, W.; Zhu, X.; Gong, S. Harmonious attention network for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2285–2294.
38. Chen, B.; Deng, W.; Hu, J. Mixed high-order attention network for person re-identification. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 371–381.
39. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6450–6458.
40. Yang, F.; Yan, K.; Lu, S.; Jia, H.; Xie, X.; Gao, W. Attention driven person re-identification. *Pattern Recognit.* **2019**, *86*, 143–155. [CrossRef]
41. Zhang, Y.; Li, K.; Li, K.; Zhong, B.; Fu, Y. Residual non-local attention networks for image restoration. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
42. Wang, X.; Girshick, R.; Gupta, A.; He, K. Non-local neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7794–7803.
43. Hinton, G.E.; Vinyals, O.; Dean, J. Distilling the knowledge in a neural network. *arXiv* **2015**, arXiv:1503.02531.
44. Romero, A.; Ballas, N.; Kahou, S.E.; Chassang, A.; Gatta, C.; Bengio, Y. Fitnets: Hints for thin deep nets. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
45. Yim, J.; Joo, D.; Bae, J.; Kim, J. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7130–7138.
46. Wang, G.; Gong, S.; Cheng, J.; Hou, Z. Faster person re-identification. In Proceedings of the European Conference on Computer Vision, Online Platform, 23–28 August 2020.
47. Zhang, Y.; Xiang, T.; Hospedales, T.M.; Lu, H. Deep mutual learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4320–4328.
48. Ding, L.; Zhou, J.; Meng, L.; Long, Z. A practical cross-view image matching method between UAV and Satellite for UAV-based geo-localization. *Remote Sens.* **2020**, *13*, 47. [CrossRef]
49. 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.
50. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
51. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
52. Radenovic, F.; Tolias, G.; Chum, O. Fine-tuning CNN image retrieval with no human annotation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *41*, 1655–1668. [CrossRef] [PubMed]