remote sensing

MDPI

Article

Remote Sensing Image Retrieval with Gabor-CA-ResNet and Split-Based Deep Feature Transform Network

Zheng Zhuo and Zhong Zhou *

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, China; zzhuo@buaa.edu.cn
* Correspondence: zz@buaa.edu.cn

Abstract: In recent years, the amount of remote sensing imagery data has increased exponentially. The ability to quickly and effectively find the required images from massive remote sensing archives is the key to the organization, management, and sharing of remote sensing image information. This paper proposes a high-resolution remote sensing image retrieval method with Gabor-CA-ResNet and a split-based deep feature transform network. The main contributions include two points. (1) For the complex texture, diverse scales, and special viewing angles of remote sensing images, A Gabor-CA-ResNet network taking ResNet as the backbone network is proposed by using Gabor to represent the spatial-frequency structure of images, channel attention (CA) mechanism to obtain stronger representative and discriminative deep features. (2) A split-based deep feature transform network is designed to divide the features extracted by the Gabor-CA-ResNet network into several segments and transform them separately for reducing the dimensionality and the storage space of deep features significantly. The experimental results on UCM, WHU-RS, RSSCN7, and AID datasets show that, compared with the state-of-the-art methods, our method can obtain competitive performance, especially for remote sensing images with rare targets and complex textures.

Keywords: high-resolution remote sensing image retrieval; Gabor; ResNet; channel attention mechanism; split

Citation: Zhuo, Z.; Zhou, Z. Remote Sensing Image Retrieval with Gabor-CA-ResNet and Split-Based Deep Feature Transform Network. *Remote Sens.* **2021**, *13*, 869. https://doi.org/10.3390/rs13050869

Academic Editors: Vasileios Syrris, Sveinung Loekken and Pedro Melo-Pinto

Received: 31 December 2020
Accepted: 23 February 2021
Published: 26 February 2021

1. Introduction

In recent years, with the increasing demand for high-resolution remote sensing data in the field of applications, the quantity of remote sensing imagery data has increased exponentially, and the quality of remote sensing data is also getting higher. How to quickly and effectively find the remote sensing image that meets the needs of the image data has become is an urgent technical problem [1]. Content-based image retrieval (CBIR) [2] is a branch of computer vision that focuses on large-scale image retrieval and is widely applied in high-resolution remote sensing image retrieval (RSIR). CBIR extracts image features to characterize the content of the image, then builds a feature library making an index for each image. The query image's features are matched to the features in the database to compute the similarity between the features so that the top-N images with similar features are returned. As we all know, CBIR relies on two key technologies: feature extraction and similarity measurement, which use image features to represent the content of the image and take the image with similar features as the retrieval result. Therefore, how to extract discriminative image features is the key technology of CBIR.

The viewing angle of remote sensing images is basically looking down from overhead with a large range from a few hundred meters to nearly 10,000 meters. Since remote sensing images have small objects with various scales, multiple directions, complex and diverse backgrounds, that is relatively rich texture, it causes a great challenge to remote sensing image retrieval. Designing an effective feature extraction method based on the characteristics of remote sensing images can contribute to improving the retrieval performance

of remote sensing images. Therefore, we propose the Gabor-CA-ResNet and split-based deep feature transform network. Firstly, in view of the advantages of Gabor in describing the image space-frequency structure, the Gabor filter is added to the ResNet network to enhance the discriminative ability of deep features in texture, orientation, and scale changes. Then, the channel attention mechanism is introduced to extract more representative and discriminative image features. Next, to reduce the dimensionality and the storage space of deep features, a split-based deep feature transform network is designed to transform the features from Gabor-CA-ResNet, thereby improving retrieval performance.

The paper is organized as follows. Section 2 presents and analyzes the related work summarizing the research progress in the field of remote sensing image retrieval based on deep learning. Section 3 describes the proposed Gabor-CA-ResNet network. Section 4 introduces the split-based deep feature transform network. Section 5 includes the similarity measurement method used in this paper. Section 6 describes the experimental results and analysis. Finally, Section 8 concludes the paper.

2. Related Works

In recent years, deep learning has made great breakthroughs in speech recognition, natural language processing, computer vision, and many other fields [3]. As one of the representative algorithms of deep learning, convolutional neural networks (CNN) have achieved the best results in computer vision, classification, and other fields. It adopts deep hierarchical architectures with parameters of each layer learned from large labeled classification datasets [4]. The advantage of deep features is to extract information layer by layer from pixel-level raw data to abstract semantic concepts, which makes it have outstanding advantages in extracting global features and context information of images. In recent years, researchers have applied CNN to image retrieval and achieved performance far exceeding traditional methods, and it has become the current mainstream solution of high-resolution remote sensing image retrieval.

CNN-based high-resolution remote sensing image retrieval schemes mainly include two types: image retrieval based on classification network and image retrieval based on retrieval network. The following describes the research progress of the two schemes.

2.1. Remote Sensing Image Retrieval Based on Classification Network

Deep feature extraction and similarity measurement are implemented independently in this kind of scheme. After training a classification CNN, the classification network is used to extract deep features to represent the content of the image. Unlike handcrafted features, CNN-based feature extraction is driven by data and can automatically learn feature representations from big data, but usually requires complex model parameters. The deep features are mainly extracted from the convolutional layer and fully connected layer of CNN. The convolutional layer features contain more details from low-layer CNN, and the fully connected layer features focus more on semantics from the high-layer CNN.

Napoletano et al. [5] conducted a comparative study on CNN features and handcrafted features such as LBP and SIFT, respectively, proposed four different retrieval schemes and conducted experiments on UCM and WHU-RS datasets. The results show that CNN features can obtain the best retrieval performance. Zhou et al. [6] proposed the LDCNN (Low-Dimensional CNN) by combining the CNN and the NIN (Network in Network) network to adopt two CNN strategies for remote sensing image retrieval. The first strategy is to extract features from the fully connected layer and the convolutional layer of the pre-trained CNN; the second strategy is to fine-tune the pre-trained CNN model on the target dataset or design a new network structure. The LDCNN is trained on a large-scale remote sensing dataset. The experimental results with the two strategies show that LDCNN can achieve better results. Wang et al. [7] proposed an image retrieval method based on bilinear pooling, in which the ImageNet dataset [8] was used to pre-train the VGG-16 [9] and ResNet34 [10] networks, and the convolutional layer features of the two networks was weighted through the channel and spatial attention mechanism to retrieve useful feature

channels for the task to assign higher weight. Then, the deep features are obtained by using the bilinear pooling method after extracting the last convolutional layer features of the two networks. Finally, the dimensionality of deep features for image retrieval is reduced by using PCA. The research results show that the retrieval performance is better than other pooling methods. The features of the fully connected layer mainly contain semantic information, which lacks local details and location information of the image. For this reason, Hu et al. [11] and Xia et al. [12] proposed a fully connected layer feature extraction method based on multiple blocks or regions, in which the image was firstly divided into blocks to extract the fully connected layer features of each block separately, and cascades them, and then the maximum pooling, average pooling, and hybrid pooling methods were used to aggregate these features, and finally PCA dimensionality reduction was used to obtain low-dimensional features. The results show that extracting fully connected layer features by the block can solve the problem that the fully connected layer features cannot provide location information. Compared with the method of extracting fully connected layer features from the entire image, it can effectively improve retrieval performance.

Recently, some works have been attempted after researchers recognized the necessity of improving model conversion capabilities. Dai et al. [13] introduced a deformable convolution filter to enhance the geometric transformation modeling ability of CNN. It allows free deformation of the sampling grid, and its offset is learned from the previous feature map. However, deformable filtering is more complicated and is related to the region of interest (RoI) pooling technology originally designed for target detection [14]. Zhou et al. [15] proposed an active rotation filter to make CNN have the generalization ability. However, this filter rotation method is actually only suitable for small and simple filters. Jacobsen et al. [16] proved that regularization on the filter function space can improve generalization ability by combining low-order filters with learned weight coefficients, but it is only available for training small datasets.

2.2. Remote Sensing Image Retrieval Based on Retrieval Networks

Different from the above retrieval scheme, the retrieval scheme based on a retrieval network is realized by designing a special retrieval CNN, and integrating feature extraction and similarity measurement into a unified framework, while extracting image features, so as to reduce the distance between similar images, and enlarge the distance between dissimilar images.

Some researchers have devoted themselves to remote sensing image retrieval based on retrieval networks and achieved good results. Ye et al. [17] took the advantage of the similarity between image categories and first obtained the initial retrieval results by sorting the CNN feature distance between the query image and each retrieved image in ascending order. Then, the weight between the query image and each class is calculated according to the initial results, here, the initial retrieval results are reordered, and the retrieval performance obtained is better than the state-of-the-art method. Cao et al. [18] proposed a three-tuple network, which outputs feature vectors of images, positive and negative samples, and normalizes them. The loss value is calculated by the distance between feature vectors, and the distance between positive samples is got closer, while the distance between negative samples is pushed forward. The final retrieval performance is significantly better than the existing methods. Zhang et al. [19] proposed a hyperspectral remote sensing image retrieval scheme by using unsupervised learning to train the DCGAN network, extracting features for retrieval, and introducing relevance feedback based on feature weighting to further improve retrieval accuracy. Keisler [1] used an autoencoder to compress the 2048 features from the penultimate layer of ResNet50 into 512 binary features, which is able to search over approximately 2 billion images in 0.1 s due to using a hash-based search method.

Based on the above analysis, image retrieval focuses on feature extraction and ditance measurement. The more discriminative the extracted features are, the better the retrieval performance will be. This paper focuses on feature extraction. We propose a remote

sensing image retrieval method with Gabor-CA-ResNet and a split-based deep feature transform network. The framework of the proposed method is shown in Figure 1. Firstly, a Gabor-CA-ResNet network is proposed to extract the deep features of images; then, a split-based deep feature transform network is designed to reduce the dimensionality while improving the discriminative ability of features; finally, L2 distance is used to measure the similarity to realize remote sensing image retrieval.

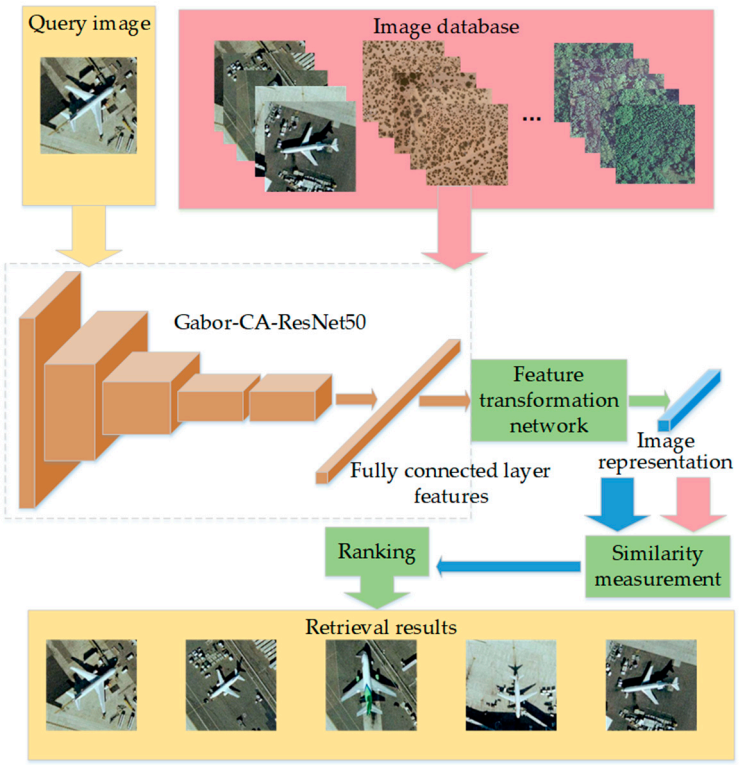

Figure 1. The remote sensing image retrieval framework of Gabor-CA-ResNet and split-based deep feature transform network.

3. Proposed Gabor-CA-ResNet Network

We use ResNet as the backbone network proposed by He et al. [10] in 2015. The author proposes a deep residual learning framework to solve the problem of performance degradation due to the increase of depth. By superimposing identity mapping on a shallow network, the network will not degenerate with the increase of depth. In recent years, ResNet has been widely cited in various computer vision tasks and has achieved outstanding performance. However, ResNet is initially designed for natural images, if being applied to remote sensing image processing, it needs to consider the characteristics of remote sensing images to design corresponding mechanisms. Therefore, we propose a Gabor-CA-ResNet network structure based on the characteristics of remote sensing images with rich texture, different object scales, and multiple orientations, as shown in Figure 2. Considering that Gabor is mainly used to enhance the ability of deep features for representing texture, direction, and scale changes, a Gabor convolutional layer is added to the lower layer of ResNet. Moreover, to obtain semantic features, we introduce a channel attention mechanism to the high layer of ResNet to further enhance the discriminative ability of features.

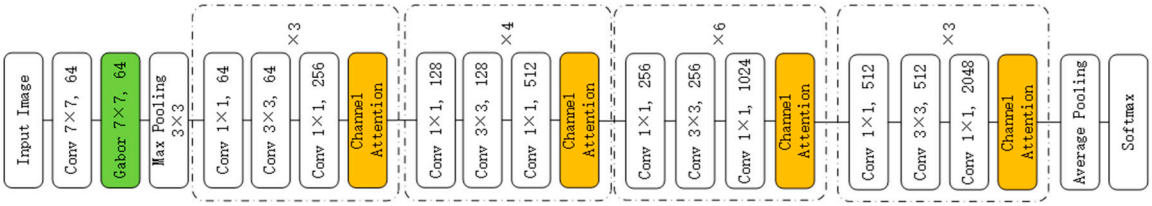

Figure 2. Gabor-CA-ResNet network structure.

The following describes the implementation details of the Gabor convolutional layer and channel attention mechanism.

After the input image passes through the general convolutional layer once, the low-level features are enhanced through the controllable Gabor convolutional layer and then sent to the residual network, in which each residual block follows a channel attention module to further enhance the features. Finally, after the high-level features are processed through the average pooling layer, the classification results of the image are output through the Softmax layer, here, the cross-entropy function is used to calculate the loss.

3.1. Gabor Convolutional Layer

Anisotropic filtering technology is widely used to extract robust image features, in which Gabor is the most representative filter among them. The Gabor transform uses a set of Gabor filters with different time-frequency domain characteristics as the basis function for image transformation. Each channel can obtain a certain local feature of the input image to describe the space in the image while preserving the spatial relationship information–frequency structure, with multi-resolution characteristics.

It can be seen that Gabor is very similar to the visual stimulus response of simple cells in the human visual system. This makes the Gabor transform has significant advantages in extracting the local space-frequency domain features of the target, so it is widely used in image processing, pattern recognition, and other fields.

The Gabor wavelet transform is defined as follows:

$$g(x, y; \lambda, \theta, \Psi, \sigma, \gamma) = \exp\left(-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right) \exp\left(i\left(2\pi\frac{x'}{\lambda} + \Psi\right)\right) \quad (1)$$

among them,

$$x' = x\cos\theta + y\sin\theta$$

$$y' = y\cos\theta - x\sin\theta$$

where λ is the wavelength of the cosine function, θ is the direction of the parallel fringes between the normal and the Gabor kernel function, ψ is the phase shift, σ is the standard deviation of the Gaussian factor of the Gabor kernel function, and γ is the spatial aspect ratio that determines the shape of the Gabor kernel function.

Here, we add a Gabor convolutional layer to ResNet with a controllable convolutional layer shown in Figure 3. The weight of the controlled convolutional layer comes from Equation (1), and the weight of each channel is generated by different Gabor parameters. In this paper, the Gabor layer contains 64 channels.

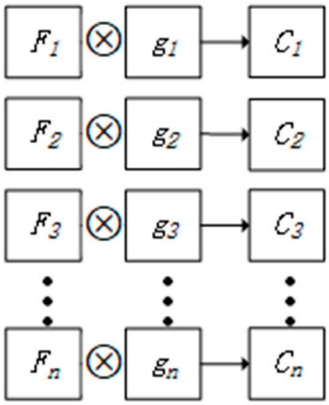

Figure 3. Gabor convolutional layer structure.

In Figure 3, assuming F_i (i = 1, 2 . . . n) represent the ith channel of the input, g_i (i = 1, 2 . . . n) is the Gabor kernel function, then the output feature channel C_i (i = 1, 2 . . . n) of the Gabor convolutional layer can be expressed as:

$$C_i = F_i \otimes g_i \tag{2}$$

The Gabor kernel function has many parameters. Through a large number of experiments, we found that the λ, ψ, σ, and γ parameters have little effect on the retrieval performance, while the parameter θ has a direct effect on the retrieval performance. In order not to increase the number of training parameters, we set the parameters λ, ψ, σ, and γ in the Gabor kernel function to be fixed and not participate in the iteration. Only the cases where θ is 0°, 45°, 90°, and 135° are considered.

In order to allow the Gabor convolutional layer to adapt to any convolutional network, we set the size of the convolution kernel to $2n$+1 (n is a natural number), the convolution step size to 1, and the Gabor convolutional layer is input by zero paddings, the size of the output feature map is kept consistent so that it is convenient to insert the Gabor convolutional layer into any position of the original CNN.

3.2. Channel Attention Mechanism

Attention mechanism comes from the study of human vision. Due to the bottleneck of information processing, human beings selectively pay attention to a part of all information and ignore other visible information. The channel attention mechanism improves the representation ability of the network by modeling the dependence of each channel and can adjust the features channel by channel so that the network can learn to selectively strengthen the features containing useful information and suppress useless features through global information [20].

The principle of the channel attention mechanism is shown in Figure 4, which is divided into three parts: squeeze, excitation, and scale. Firstly, the output signal of each channel is considered, the global spatial information is compressed into channel descriptors, and the global average pooling is used to generate the statistics of each channel. It can be expressed by the mathematical formula:

$$z_c = \mathrm{F}_{\mathrm{squeeze}}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{3}$$

where C represents the Cth convolution kernel in the convolution layer, and H and W represent the size of the convolution kernel.

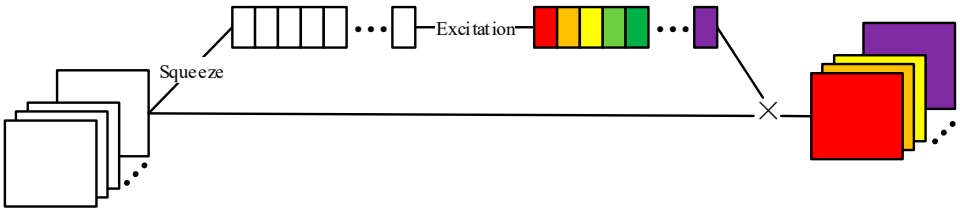

Figure 4. Principle of channel attention mechanism.

The second is to investigate the dependence of each channel, because multiple channels may affect the results. In this paper, we use the threshold mechanism with sigmoid activation. To limit the complexity of the model and enhance the generalization ability, two fully connected layers in the form of a bottleneck are used in the threshold mechanism. The dimensionality of the first fully connected layer is reduced to 1/r, and r is a hyper-parameter. The incentive mechanism in the form of Sigmoid is adopted as:

$$\mathrm{s} = \mathrm{F}_{\mathrm{excitation}}(z, W) = \sigma(g(z, W)) = \sigma(W_2 \mathrm{Relu}(W_1 z)) \tag{4}$$

where $W_1 \in R^{\frac{C}{r} \times C}$, $W_2 \in R^{C \times \frac{C}{r}}$. The final sigmoid function is the weight of each channel. Adjusting the weight of each channel feature according to the input data helps to enhance the distinguishability of features. Finally, the activation value is multiplied correspondingly to the original feature channel as:

$$\widetilde{x} = \mathrm{F}_{\mathrm{scale}}(u_c, s_c) = s_c \cdot u_c \tag{5}$$

Through a lot of experiments, Hu et al. [20] conducted experiments for a range of different r values, and the conclusion is that when r = 16, accuracy and complexity can be well balanced. The channel attention mechanism dynamically adjusts the characteristics of each channel according to the input to enhance the representation ability of the network. In addition, it can also be used to assist network compression.

3.3. Network Training Details

We use the "pre-training + fine-tuning" strategy to train the network. First, the ImageNet dataset is used to pre-train the ResNet network to obtain the network model parameters. The features extracted by the model pre-trained with the ImageNet dataset have the strong distinguishing ability. Using a remote sensing image dataset to fine-tune the model, better performance can be obtained. This method is widely used in image retrieval. Next, the remote sensing image dataset is used to fine-tune the Gabor-CA-ResNet network, in which the initial parameters of the ResNet part change to the pre-trained ResNet model parameters. Through fine-tuning, the optimized Gabor-CA-ResNet network model is obtained. The output of the last pooling layer of the Gabor-CA-ResNet network is extracted as the deep feature of the image, and its dimension is 2048.

4. Split-Based Deep Feature Transform Network

The deep feature dimensionality extracted by the Gabor-CA-ResNet network is 2048. Such a high dimensionality will bring a lot of pressure and burden to the subsequent calculation and storage. To this end, we design a split-based deep feature transform network reducing the dimensionality and enhancing the discriminative ability of features.

The overall structure of the split-based deep feature transform network designed is shown in Figure 5. This structure does not directly transform the input N-dimensional Gabor-CA-ResNet deep features, but first divides it into M segments, in which the segment length is $N_1 = N/M$. Each segment is transformed through a fully connected neural network, ($N_1 > N_2 > N_3$), so that the transformed feature has a lower dimensionality than the original feature. Each segment uses the same transform network structure, cascades the output of each segment, and outputs the classification result of the image through a Softmax, and

uses the cross-entropy function to calculate the loss. The deep features extracted by the Gabor-CA-ResNet network and the image categories form training samples for training the feature transform network. After training, the $M \times N_3$-dimensional feature transform result in Figure 5 is extracted as the image feature. L2 distance measurement criterion is used for similarity computation to realize image retrieval. Experimental results show that, compared with non-segmentation, the use of such a segmented structure design can not only greatly reduce the calculation amount of the feature transform network while improving the training efficiency, but also obtain better retrieval performance.

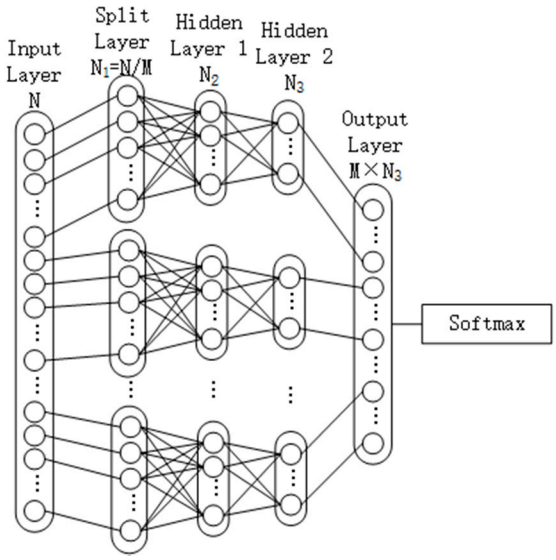

Figure 5. The structure of split-based deep feature transform network.

5. Similarity Measurement

The Euclidean distance of feature vectors is adopted to measure the similarity of images. The Euclidean distance between two points x_{1k} ($k = 1, 2 \ldots n$) and x_{2k} ($k = 1, 2 \ldots n$) in N-dimensional space is defined as follows:

$$d_{12} = \sqrt{\sum_{k=1}^{n} (x_{1k} - x_{2k})^2} \tag{6}$$

Euclidean distance is widely used in image retrieval similarity measurement, and it is one of the most effective and widely used measurement methods.

6. Experimental Results and Analysis

In order to evaluate the performance of the proposed method, we have made comparison experiments on four high-resolution remote sensing image datasets including UCM, RS19, RSSCN7, and AID. The experimental results are introduced and the results are analyzed.

6.1. Datasets and Evaluation Metric

UCM, WHU-RS, RSSCN7, and AID are the four most commonly used high-resolution remote sensing image datasets.

The images in the UCM dataset [21] come from the U.S. Geological Survey's city map, which includes 21 types of scene images, such as airplanes, beaches, buildings, and dense residential areas. Each type has 100 images of 256 × 256 size, and the spatial resolution of each pixel is 0.3 m.

The WHU-RS dataset [22] is a remote sensing image dataset released by Wuhan University in 2011. The pixel size of each image is 600 × 600, and there are 19 types of scene images, each of them contains about 50 images, with a total of 1005 images.

The RSSCN7 dataset [23] is a remote sensing image dataset released by Wuhan University in 2015, which contains 2800 images. These images come from seven typical scenes: farmland, parking lot, residential area, industrial area, and lake. Each category includes 400 images with the size of 400 × 400. Due to the diversity of scenarios, this dataset is very challenging.

AID dataset [24] is a remote sensing image dataset jointly released by Wuhan University and Huazhong University of Science and Technology in 2017. It contains 30 types of scene images, each type contains about 220–420 images, a total of 10,000 images, and the pixel size of images is 600 × 600.

We use mean average precision (mAP) [25] to evaluate retrieval performance, which is the public accepted image retrieval performance evaluation index. The mAP is defined as follows:

$$\text{mAP} = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \tag{7}$$

The definition of *AveP* is:

$$AveP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{number\ of\ relevant\ images} \tag{8}$$

where Q is the number of all images in the dataset, $P(k)$ is the accuracy rate and $rel(k)$ is a piecewise function. The precision is calculated once for every image returned and multiplied by the precision by the coefficient $rel(k)$. If the current returned image is related, the $rel(k)$ is 1; otherwise, it is 0.

6.2. Experimental Setting

In our experiments, the dataset is randomly divided into two parts, and experiments were repeated five times, with the final results being the average. In the experiment, 80% of the images are training samples that are randomly selected from each dataset, and the remaining 20% of the images are used as test samples. The training samples were expanded 16 times by the rotating original image and its horizontal mirror image once every 45 ° and all the images are normalized to 224 × 224 pixels.

Our network is built and tested in the Keras open source framework. The experimental platform uses Intel Core i7-8700, CPU 3.2GHz, 32GB memory, 4T hard disk space including NVIDIA GeForce RTX 2080 Ti graphics card for training and testing. The learning rate is 0.01, the batch size is set to 16, and the number of training iterations is set to 50 rounds. The momentum attenuation method and weight attenuation method are used to optimize the training process to prevent overfitting. The weight decay rate is 0.0001, and the momentum parameter is set to 0.9.

6.3. Experiment I: Performance Comparison of Deep Feature Extraction Networks

In order to verify the feature discriminative ability of our Gabor-CA-ResNet network, we compare to ResNet50 and the ResNet50 network that introduced the Gabor convolutional layer. In the experiment, the feature dimensionality extracted by each network is 2048, and the L2 distance measurement criterion is used for similarity comparison. The experimental results on the four datasets are shown in Table 1.

Table 1. Performance comparison of mean Average Precision (mAP) with different network structures.

Structure / Dataset	UCM	WHU-RS	RSSCN7	AID
ResNet	90.63%	93.76%	74.22%	81.37%
Gabor-ResNet	94.78%	97.37%	77.20%	87.69%
Gabor-CA-ResNet	97.50%	99.48%	85.96%	90.52%

It can be seen from Table 1 that for different datasets, Gabor-ResNet introduces Gabor filters, and the retrieval performance is 94.78%, 97.37%, 77.20%, and 87.69%, respectively. Compared with the ResNet network, they are improved by 4.15%, 3.61%, 2.98%, and 6.32%. This shows that Gabor can effectively enhance the discriminative ability of deep features in terms of texture, direction, and scale changes. After adding the channel attention mechanism, the retrieval performance has been further improved, reaching 97.50%, 99.48%, 85.96%, and 90.52% on the four datasets, respectively, and the increase rates are 2.72%, 2.11%, 8.76%, and 2.83%, respectively. This presents that the channel attention mechanism can obtain more representative and discriminative features. The deep features extracted by the Gabor-CA-ResNet network have a strong representative and discriminative abilities, especially in the RSSCN7 dataset with fewer targets and rich textures. The increase was the largest, reaching 11.74%.

6.4. Experiment II: The Impact of Split-Based Deep Feature Transform Network Structure on Performance

The feature transform network includes more parameters. In order to verify the influence of different parameters on the retrieval performance, we set different parameters and conducted experiments on the three datasets of UCM, WHU-RS, and RSSCN7. The number of segments M is set to 4, 8, and 16, respectively; N_2 is set to 16, 32, 64, 128, and N_3 is set to 2, 4, 8, 16, respectively. When setting parameters, ensure that $N_1 > N_2 > N_3$, so that the output feature has a lower dimensionality than the original feature. The transformed feature dimensionality is $M \times N_3$. After many experimental verifications, and comprehensively considering the feature dimensionality and retrieval performance after dimensionality reduction, when the parameters are set to $M = 8$, $N_2 = 64$, and $N_3 = 8$, the best performance feature can be obtained.

We compared the unsplit deep feature transform network with the proposed split-based deep feature transform network in terms of retrieval performance and the number of parameters. The results are shown in Table 2. In the table, DNN 256-64 represents the unsplit deep feature transform network, the input dimensionality is 2048, the number of neurons in the two fully connected layers is 256 and 64, respectively, DNN 8-64-8 represents the split-based deep feature transform network, network parameters Set as $M = 8$, $N_2 = 64$, $N_3 = 8$, the best retrieval performance can be obtained.

Table 2. Comparison of the impact of feature transform network structure on performance.

Structure	DNN 256-64	DNN 8-64-8	Origin
Dimensionality	64	64	2048
Parameters	542,357	137,109	-
UCM	98.47%	98.68%	97.50%
RS19	99.59%	99.70%	99.48%
RSSCN7	95.17%	96.61%	85.96%
AID	94.34%	95.75%	90.52%

It can be seen from the table that for all datasets, the features after the feature transform network can obtain higher retrieval performance than the original features, which shows that the feature transform network designed in this paper can effectively improve the discriminative ability of features. Moreover, compared with the unsplit feature transformation, the split-based deep feature transform network can reduce network complexity

while improving retrieval performance. The number of parameters of the split-based deep feature transform network is only 25.28% of that of the unsplit network. However, in terms of retrieval performance, UCM, the four datasets of RS19, RSSCN7, and AID reached 98.68%, 99.70%, 96.61% and 95.75%, which increased by 0.21%, 0.11%, 1.44% and 1.41%, respectively. The precision of the split network is similar to that of the unsplit network, but the parameters are greatly reduced so that a larger batch size can be set and the training speed is faster.

6.5. Experiment III: Comparison of Dimensionality Reduction Performance of Feature Transform Networks

The feature transform network can not only improve the retrieval performance but also reduce the feature dimensions, retrieval time, and storage space. We reduce the feature dimensions from 2048 to 64. In order to further verify the dimensionality reduction performance of the feature transform network, we also compare to the common dimensionality reduction methods such as PCA [26], LPP [27], and Isomap [28]. PCA is the most representative linear dimensionality reduction method, and LPP is a representative non-linear dimensionality reduction method. The main advantage of Isomap is to use "geometric distance" instead of the original Euclidean distance. In this way, the loss of data information can be better controlled, and the data in the high-dimensional space can be more comprehensively displayed in the low-dimensional space. The target dimensionality of the three dimensionality reduction methods is all set to 64. The experimental results of the feature transform network and other dimensionality reduction methods are shown in Table 3.

Table 3. Performance comparison of feature transform network and other dimensionality reduction methods.

Method	UCM	WHU-RS	RSSCN7	AID
Original Feature	97.50%	99.48%	85.96%	90.52%
PCA	97.61%	99.55%	86.61%	90.54%
LPP	95.58%	99.28%	80.47%	82.64%
Isomap	92.15%	98.55%	80.56%	74.49%
Ours	98.68%	99.70%	96.61%	95.75%

It can be clearly seen from Table 3 that for the four datasets, compared with other dimensionality reduction methods, the feature transform network can obtain better retrieval performance. PCA, LPP, and Isomap are all unsupervised dimensionality reduction methods. Feature transform network is a supervised dimensionality reduction method, which can further improve the discriminative ability of features while reducing dimensionality, and achieve better retrieval performance.

6.6. Experiment IV: Performance Comparison of Euclidean and Other Similarity Measurement Methods

To verify the performance of the Euclidean distance, we also compared it with the other classical similarity measurement methods such as Cityblock, Chebychev, Cosine, Correlation, and Spearman. The experimental comparison results of Euclidean distance and other similarity measurement methods are shown in Table 4.

Table 4. Performance comparison of Euclidean distance and other similarity measurement methods.

Method	UCM	WHU-RS	RSSCN7	AID
Euclidean	98.68%	99.70%	96.61%	95.75%
Chebychev	97.38%	99.02%	93.86%	93.78%
Cosine	98.68%	99.70%	96.61%	95.75%
Correlation	98.63%	99.73%	96.60%	95.78%

It can be seen from Table 4 that the performance of Cosine and Euclidean is similar, and Correlation and Euclidean have their advantages in performance. Considering comprehensively, this solution selects Euclidean distance which has stable performance and is widely used.

6.7. Experiment V: Image Retrieval Results

The results of remote sensing image retrieval are shown in Figure 6. Figure 6 shows the first five images of the partial retrieval results of the UCM dataset. From the retrieval results, the method in this paper can obtain better retrieval results. The images with high similarity to the query images are ranked ahead in the retrieval results.

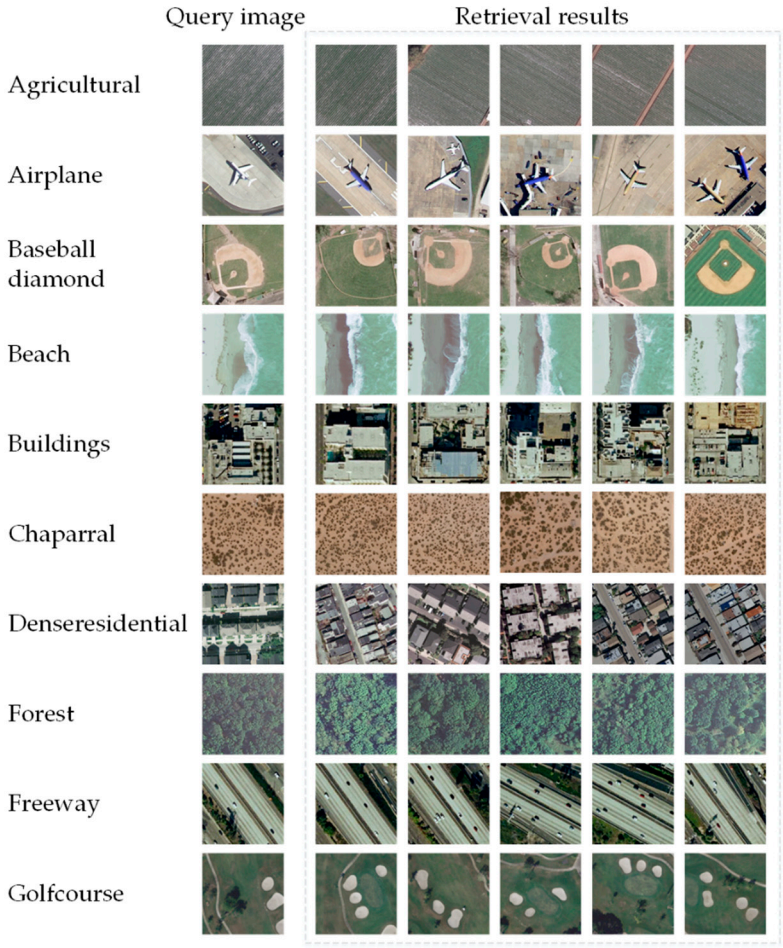

Figure 6. Top five image retrieval results of UCM dataset.

6.8. Experiment VI: Performance of Each Category

In order to show the effectiveness of our method more clearly, Table 5 shows the retrieval performance of each category of the UCM dataset. The UCM dataset contains 21 categories, including agricultural, airport, baseballdiamond, beach, buildings, chaparral, denseresidential, forest, freedom, golf course, harbor, intersection, medium resident, mobilehomepark, overpass, parkinglot, river, runway, sparsuresidential, storagetanks and tenniscourt.

Table 5. Performance of each category of the UCM dataset.

Categories	Average Precision
Agricultural	95.01%
Airplane	96.57%
Baseballdiamond	100.00%
Beach	100.00%
Buildings	99.24%
Chaparral	100.00%
Denseresidential	92.30%
Forest	100.00%
Freeway	100.00%
Golfcourse	99.99%
Harbor	100.00%
Intersection	100.00%
Mediumresidential	95.44%
Mobilehomepark	100.00%
Overpass	100.00%
Parkinglot	100.00%
River	99.71%
Runway	99.97%
Sparseresidential	94.82%
Storagetanks	99.92%
Tenniscourt	99.30%

From Table 5, we can see that our method achieves 100% accuracy in the category of complex texture, such as baseballdiamond, beach, chaparral, forest, freedom, etc. For complex objects, the average precision is low, such as densersidentity, the accuracy is 92.30%. The feature extraction network we designed is improved from the classification network. The images of the densersidentity category are similar to the buildings category, which leads to the extracted features biased to the wrong category.

6.9. Experiment VII: Comparison with Existing Methods

In order to verify the effectiveness of our method, we compare it with the existing five high-resolution RSIR methods. The retrieval performance comparison results on four datasets are shown in Table 6, in which the results of the other methods are referenced from the literature.

Table 6. Performance comparison with existing methods.

Author	Year	UCM	WHU-RS	RSSCN7	AID
Napoletano P. [5]	2017	98.05%	98.69%	-	-
Zhou W. [6]	2017	54.44%	64.60%	46.28%	37.61%
Li Y. [29]	2018	70.39%	-	-	-
Wang Y. [7]	2019	90.56%	89.51%	81.32%	-
Ye F. [17]	2019	95.62%	-	-	-
Ours		98.68%	99.70%	96.61%	95.75%

As can be seen from Table 6, our method can obtain the state-of-the-art retrieval performance on all four datasets compared with the existing method. For the most challenging RSSCN7 dataset, our method can also achieve mAP of 96.61%. On the WHU-RS dataset, it has reached 99.70%. This is because this solution introduces the Gabor convolutional layer and channel attention mechanism in the ResNet network according to the characteristics of remote sensing images, which can extract more representative and discriminative image features. The use of a split-based deep feature transform network can not only reduce the computational complexity but also further improve the discriminative ability of deep features, thereby obtaining the best retrieval performance.

In order to further demonstrate the effectiveness of this method, Figure 7 shows the precision-recall curve of different methods on the UCM dataset. The curves of other methods are from the literature [5,29,30]. It can be seen from Figure 7 that the PR curve of this method is higher than other methods, and the results prove that the retrieval effect of this method is better than that of existing methods.

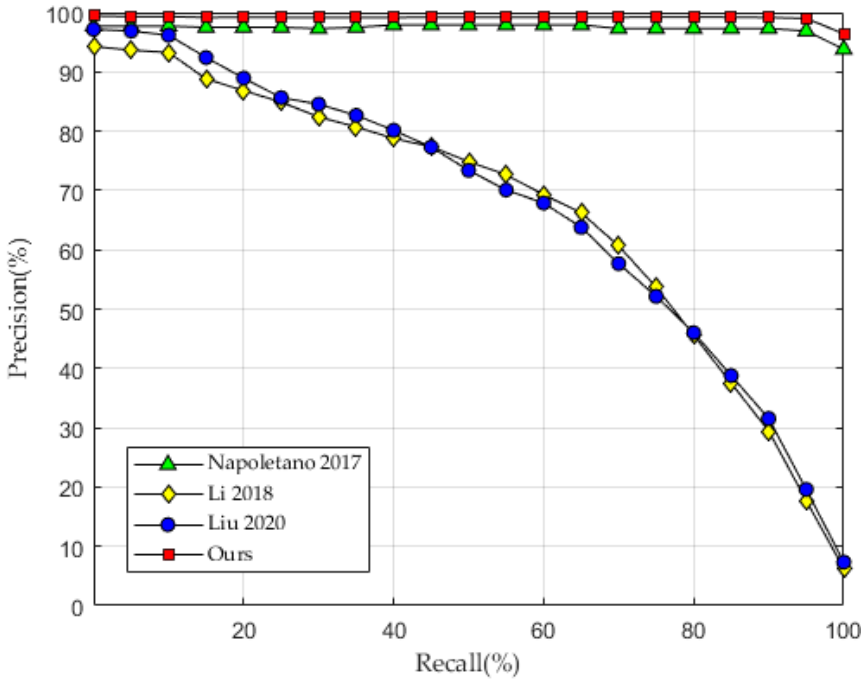

Figure 7. Comparison of PR curves between our method and other methods.

7. Discussion

In this paper, the Gabor filter is introduced into ResNet to enhance the ability of deep learning features to describe the texture, direction, and scale changes. Furthermore, we add the channel attention mechanism to enhance the discriminative ability of important characteristic channels. At the same time, we design a split-based deep feature transform network to improve the feature discriminative ability while reducing the dimensionality of features, thereby improving retrieval performance and greatly reducing storage space. The number of parameters of the split-based deep feature transform network is only 25.28% of that of the unsplit network, which can greatly reduce the demand for computing resources and improve the training speed.

Experiments verify the effectiveness of the proposed method. Experiment I compares the performance of the proposed Gabor-CA-ResNet network structure with ResNet and Gabor-ResNet network structures on the four datasets of UCM, WHU-RS, RSSCN7, and AID. Experimental results show that Gabor-CA-ResNet can significantly improve retrieval performance. Experiment II discusses the influence of the feature transform network parameters on performance. We use different feature transform network parameters to conduct verification experiments on 4 datasets. The results show that the feature transform network effectively improved the discriminative ability of features, thereby improving retrieval performance. Experiment III compared the split-based deep feature transform net-work with PCA, LPP, Isomap, and other feature dimensionality reduction methods. The experimental results show that the split-based deep feature transform network designed in this paper can play a role in feature dimensionality reduction, and can obtain better retrieval performance. In Experiment IV, we verified the performance of the Euclidean distance. Compared with the other classical similarity measurement methods such as Chebychev, Cosine, and Correlation, the Euclidean distance obtained better optimal in most cases. In Experiment V, the top five images from our retrieval method were shown.

Our method obtained the best retrieval results. The images with high similarity to the query image ranked ahead in our retrieval results. In Experiment VI, the performance of each category was shown. Our method achieves 100% accuracy in the category of complex texture. For complex objects, the average precision is low. In Experiment VII, we compare with the existing five most advanced remote sensing image retrieval methods. It can be seen that our method obtains competitive retrieval performance.

In summary, the Gabor-CA-ResNet network can obtain deep features with the strong discriminative ability and good discrimination, split-based deep feature transform network reduces the dimensionality of the features, save storage space, so as to further improve the discriminative ability of deep features to obtain better retrieval performance.

8. Conclusions

In this paper, a Gabor-CA-ResNet network is proposed for extracting features of remote sensing images, aiming at the characteristics of high-resolution remote sensing images, and a split-based deep feature transform network is designed to further improve the discriminative ability of features, and greatly reducing storage space. We evaluate the proposed method and other retrieval methods on four high-resolution remote sensing image datasets. Experimental results show that our method can obtain competitive retrieval performance. It can be seen from the experiments that image retrieval still needs to be improved. In the next work, we will try to design a new network to extract more representative and discriminative deep features to obtain higher retrieval performance, and test cross source retrieval to verify the migration ability of the new method.

Author Contributions: Z.Z. (Zhong Zhou) and Z.Z. (Zheng Zhuo) conceived and designed the experiments, Z.Z. (Zhong Zhou) analyzed and interpreted the data and wrote the paper. Z.Z. (Zheng Zhuo) supervised the study and reviewed this paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key R&D Program of China under Grant, grant number 2018YFB2100601, and National Natural Science Foundation of China, grant number 61872023.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: UCM: http://weegee.vision.ucmerced.edu/datasets/landuse.html (accessed on 25 February 2021), WHU-RS: http://dsp.whu.edu.cn/cn/staff/yw/HRSscene.html (accessed on 25 February 2021), RSSCN7: https://sites.google.com/site/qinzoucn/documents (accessed on 25 February 2021), AID: https://pan.baidu.com/s/1mifOBv6#list/path=%2F (accessed on 25 February 2021). Some data, models, and code generated during the study are available online (https://github.com/buaavrlab/Gabor-CA-ResNet-and-Split-Based-Deep-Feature-Transform-Network (accessed on 25 February 2021)).

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

References

1. Keisler, R.; Skillman, S.W.; Gonnabathula, S.; Poehnelt, J.; Rudelis, X.; Warren, M.S. Visual search over billions of aerial and satellite images. *Comput. Vis. Image Und.* **2019**, *187*, 102790. [CrossRef]
2. Faloutsos, C.; Barber, R.; Flickner, M.; Hafner, J.; Niblack, W.; Petkovic, D.; Equitz, W. Efficient and effective querying by image content. *J. Intell. Inf. Syst.* **1994**, *3*, 231–262. [CrossRef]
3. Zhuo, Z.; Zhou, Z. Low dimensional discriminative representation of fully connected layer features using extended largevis method for high-resolution remote sensing image retrieval. *Sensors* **2020**, *20*, 4718. [CrossRef] [PubMed]
4. Krizhevsky, A.; Sutskever, I.; Hinton, G. Imagenet classification with deep convolutional neural networks. In Proceedings of the Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
5. Napoletano, P. Visual descriptors for content-based retrieval of remote-sensing images. *Int. J. Remote Sens.* **2018**, *39*, 1343–1376. [CrossRef]
6. Zhou, W.; Newsam, S.; Li, C.; Shao, Z. Learning low dimensional convolutional neural networks for high-resolution remote sensing image retrieval. *Remote Sens.* **2017**, *9*, 489. [CrossRef]
7. Wang, Y.; Ji, S.; Lu, M.; Zhang, Y. Attention boosted bilinear pooling for remote sensing image retrieval. *Int. J. Remote Sens.* **2020**, *41*, 2704–2724. [CrossRef]

8. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Li, F. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Miami Beach, FL, USA, 20–25 June 2009; pp. 248–255.
9. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
10. 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, Seattle, WA, USA, 27–30 June 2016; pp. 770–778.
11. Hu, F.; Tong, X.; Xia, G.; Zhang, L. Delving into deep representations for remote sensing image retrieval. In Proceedings of the IEEE International Conference on Signal Processing, Chengdu, China, 6–10 November 2016; pp. 198–203.
12. Xia, G.; Tong, X.; Hu, F.; Zhong, Y.; Datcu, M.; Zhang, L. Exploiting deep features for remote sensing image retrieval—A systematic investigation. *arXiv* **2017**, arXiv:1707.07321.
13. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable convolutional networks. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22 October 2017; pp. 764–773.
14. Girshick, R. Fast r-cnn. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
15. Zhou, Y.; Ye, Q.; Qiang, Q.; Jiao, J. Oriented response networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4961–4970.
16. Jacobsen, J.; van Gemert, J.; Lou, Z.; Smeulders, A.W.M. Structured receptive fields in cnns. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2610–2619.
17. Ye, F.; Dong, M.; Luo, W.; Chen, X.; Min, W. A new re-ranking method based on convolutional neural network and two image-to-class distances for remote sensing image retrieval. *IEEE Access* **2019**, *7*, 141498–141507. [CrossRef]
18. Cao, R.; Zhang, Q.; Zhu, J.; Li, Q.; Li, Q.; Liu, B.; Qiu, G. Enhancing remote sensing image retrieval using a triplet deep metric learning network. *Int. J. Remote Sens.* **2020**, *41*, 740–751. [CrossRef]
19. Zhang, J.; Chen, L.; Zhuo, L.; Liang, X.; Li, J. An efficient hyperspectral image retrieval method: Deep spectral-spatial feature extraction with dcgan and dimensionality reduction using t-sne-based nm hashing. *Remote Sens.* **2018**, *10*, 271. [CrossRef]
20. Hu, J.; Shen, L.; Sun, G.B.I. 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.
21. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the ACM Sigspatial International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 3–5 November 2010; pp. 270–279.
22. Dai, D.; Yang, W. Satellite image classification via two-layer sparse coding with biased image representation. *IEEE Geosci. Remote Sens.* **2011**, *8*, 173–176. [CrossRef]
23. Zou, Q.; Ni, L.; Zhang, T.; Wang, Q. Deep learning based feature selection for remote sensing scene classification. *IEEE Geosci. Remote Sens.* **2015**, *12*, 2321–2325. [CrossRef]
24. Xia, G.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. Aid: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
25. Deselaers, T.; Deselaers, T.; Keysers, D.; Keysers, D.; Ney, H.; Ney, H. Features for image retrieval: An experimental comparison. *Inf. Retr. J.* **2008**, *11*, 77–107. [CrossRef]
26. Jolliffe, I.T.; SpringerLink, O.S. *Principal Component Analysis*; Springer: New York, NY, USA, 1986.
27. He, X.; Niyogi, P. Locality preserving projections. In Proceedings of the Neural Information Processing Systems, Vancouver and Whistler, Vancouver, BC, Canada, 8–13 December 2003; pp. 234–241.
28. Tenenbaum, J.B. A global geometric framework for nonlinear dimensionality reduction. *Science* **2000**, *290*, 2319–2323. [CrossRef] [PubMed]
29. Li, Y.; Zhang, Y.; Huang, X.; Zhu, H.; Ma, J. Large-scale remote sensing image retrieval by deep hashing neural networks. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 950–965. [CrossRef]
30. Liu, Y.; Liu, Y.; Chen, C.; Ding, L. Remote-sensing image retrieval with tree-triplet-classification networks. *Neurocomputing* **2020**, *405*, 48–61. [CrossRef]