sensors

MDPI

Article

Duplex-Hierarchy Representation Learning for Remote Sensing Image Classification

Xiaobin Yuan [1,2], Jingping Zhu [1], Hao Lei [3,4,*], Shengjun Peng [5], Weidong Wang [6] and Xiaobin Li [7]

[1] The School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China
[2] The Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China
[3] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Xi'an Jiaotong University, Xi'an 710049, China
[4] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China
[5] The State Key Laboratory of Astronautic Dynamics, China Xi'an Satellite Control Center, Xi'an 710043, China
[6] PLA 63768, Xi'an 710600, China
[7] The Beijing Institute of Remote Sensing Information, Beijing 100192, China
* Correspondence: leihao.ai@xjtu.edu.cn

Abstract: Remote sensing image classification (RSIC) is designed to assign specific semantic labels to aerial images, which is significant and fundamental in many applications. In recent years, substantial work has been conducted on RSIC with the help of deep learning models. Even though these models have greatly enhanced the performance of RSIC, the issues of diversity in the same class and similarity between different classes in remote sensing images remain huge challenges for RSIC. To solve these problems, a duplex-hierarchy representation learning (DHRL) method is proposed. The proposed DHRL method aims to explore duplex-hierarchy spaces, including a common space and a label space, to learn discriminative representations for RSIC. The proposed DHRL method consists of three main steps: First, paired images are fed to a pretrained ResNet network for extracting the corresponding features. Second, the extracted features are further explored and mapped into a common space for reducing the intra-class scatter and enlarging the inter-class separation. Third, the obtained representations are used to predict the categories of the input images, and the discrimination loss in the label space is minimized to further promote the learning of discriminative representations. Meanwhile, a confusion score is computed and added to the classification loss for guiding the discriminative representation learning via backpropagation. The comprehensive experimental results show that the proposed method is superior to the existing state-of-the-art methods on two challenging remote sensing image scene datasets, demonstrating that the proposed method is significantly effective.

Keywords: remote sensing image classification; duplex hierarchy; discriminative representation; confusion score

Citation: Yuan, X.; Zhu, J.; Lei, H.; Peng, S.; Wang, W.; Li, X. Duplex-Hierarchy Representation Learning for Remote Sensing Image Classification. *Sensors* **2024**, *24*, 1130. https://doi.org/10.3390/s24041130

Academic Editor: Jan Cornelis

Received: 15 November 2023
Revised: 3 February 2024
Accepted: 4 February 2024
Published: 9 February 2024

1. Introduction

Remote sensing image classification (RSIC) allocates precise semantic descriptors to aerial images. This task holds significant importance in practical applications such as natural disaster detection [1], environmental monitoring [2], and urban planning [3]. However, RSIC still faces a great challenge: large dissimilarities in the same class and small dissimilarities between different classes. Remote sensing images may contain complex structures of abundant ground objects, representing a challenge characterized by substantial intra-class dissimilarities and limited inter-class disparities. Specifically, images of the same scene may appear to be different from each other due to the complex structures of the ground objects. In like manner, images of different scenes may appear to be similar, as they may contain common ground objects or share similar semantic information. Therefore, a discriminative feature is vital to RSIC.

Early RSIC methods [4–14] exploited handcrafted features to describe remote sensing images, such as color histograms (CHs) [4], scale-invariant feature transformation (SIFT) [9], and gray-level co-occurrence matrices (GLCMs) [7]. However, the handcrafted feature methods cannot meet the practical application requirements due to their inadequate extraction of high-level semantic information. Furthermore, these methods are limited by the amount of time and effort that they consume.

With the growth of the current deep learning domain, CNNs have achieved superior success in the field of remote sensing image classification. Compared with traditional methods, CNNs are able to extract representative features and show promising performance. Penatti et al. [15] introduced CNNs to remote sensing image classification. Maggiori et al. [16] devised an end-to-end framework for satellite imagery classification with CNNs. Liu et al. [17] proposed a multiscale CNN method to solve the scale variation of the objects in remote sensing images. In order to allow images to be input at arbitrary sizes, Xie et al. [18] designed a scale-free CNN (SF-CNN). Castelluccio et al. [19] used pretrained networks to carry out a remote sensing scene classification task and proved that CNNs always provide excellent performance.

However, these CNN-based RSIC methods face an unsatisfactory classification problem: large dissimilarities in the same class and small dissimilarities between different classes. Regarding dissimilarity in the same class, the primary hurdle stems from the large variation in features appearing in the same semantic class. Images commonly differ in terms of style, shape, size, and distribution, rendering accurate scene image classification a demanding task. Several scene images from the NWPU-RESISC45 dataset [20] are shown in Figure 1. In Figure 1a, the railway stations have different shapes, and the churches present different architectural styles. The challenge of inter-class similarity is mainly due to the existence of the same objects between different scene classes or high semantic overlapping in scene classes. For instance, as shown in Figure 1b, the scene classes of both the airport and the meadow contain the same object, namely, grass, and the tennis court and basketball court contain similar semantic information.

Based on the above challenges, a duplex-hierarchy method is proposed for RSIC in this study. This method preserves the discrimination among the samples from different semantic categories for pairs of images and further improves the classification accuracy. In pursuit of this goal, this method minimizes the discrimination loss for the samples within both the label space and the common representation space. This strategy guides the model in acquiring discriminative features. Furthermore, it simultaneously minimizes the confusion loss to guide the discriminative representation learning via backpropagation. Following the method, the label information and the classification details of image pairs are both extensively utilized to guarantee that the learned representation is highly discriminative in its semantic structure. The proposed method consists of three main steps. First, paired images are fed to a pretrained ResNet [21] network for extracting the corresponding features. Second, the extracted features are further explored and mapped into a common space. Third, the obtained representation is used to predict the class label of the input image, and the loss in the label space is minimized to further facilitate the learning of the discriminative representation. At the same time, confusion scores are calculated and added to the devised confusion loss to further improve the classification accuracy.

The key contributions of this study are outlined below:

(1) An end-to-end framework is proposed for the classification of remote sensing images, where the discriminative features are learned by measuring the differences between categories in the common space and label space simultaneously.
(2) Confusion scores between categories are calculated and embedded into the designed loss function, the confusion loss, to resolve the issues of large dissimilarities in the same class and small dissimilarities between different classes in remote sensing images by minimizing the confusion loss.

(3) A large number of experiments show that the proposed method is superior to the existing state-of-the-art methods, which proves the effectiveness of the proposed method.

Figure 1. Two major challenges degenerate the scene classification performance: (**a**) intra-class diversity: railway station (the 1st row) and church (the 2nd row) [20]; (**b**) inter-class similarity: airport vs. meadow, beach vs. desert, freeway vs. bridge, palace vs. church, and tennis court vs. basketball court (from top to bottom and from left to right) [20]. This motivates us to learn more discriminative representations so that the within-class scatter is small and the between-class separation is large.

The remainder of this paper is organized as follows: Section 2 reviews the relevant literature. Section 3 describes the proposed method in detail. Section 4 presents the experiments. Finally, Section 5 presents the conclusions.

2. Related Work

The deep network approach has gained popularity in recent years, and thus far, deep learning models have achieved excellent results in numerous computer vision applications, such as image classification [22], object recognition [23], and semantic segmentation [24], and the representation of features in images has entered a new era. In contrast to hand-crafted features, deep learning models have the capacity to acquire more robust, abstract representations and to differentiate features through deep architectural neural networks

without requiring significant engineering skills and experiential knowledge. Among these models, convolutional neural networks (CNNs) are more applicable to classifying remotely sensed image scenes and have yielded recent results [24–31]. Generally speaking, CNN-based remote sensing image scene classification can be broadly divided into three types: pretraining-based methods, fine-tuning-based methods, and retraining-based methods.

Pretraining-based methods. Pretraining-based methods use pretrained networks directly for the extraction of final features from remote sensing images. In 2015, Penatti et al. [15] proposed a method for remote sensing image classification based on CNNs, and the performance of the CNNs was better than that of low-level descriptors. Hu et al. [25] extracted feature descriptors using CNNs, and the pretrained neural network models were used for scene classification. Instead of handcrafted local features, Cheng et al. [26] used off-the-shelf CNN features to construct a convolutional feature package for remote sensing classification. To leverage semantic label information, Lu et al. [27] proposed a method of aggregating features using a CNN. Methods that employ pretrained CNNs as feature extractors are relatively straightforward, feasible, and effective on small datasets.

Fine-tuning-based methods. These methods use fine-tuned CNNs to better extract final features from scene images, and the fully trained results are better. Liu et al. [28] coupled CNNs with a hierarchical Wasserstein loss function (HW-CNNs) to improve the discrimination ability of the CNNs. Wang et al. [29] designed an ARCNet (attentional recursive convolutional network) that could highlight critical regions and ignore non-critical regions by introducing an attentional mechanism in CNNs. Although fine-tuned CNNs can obtain better results, they are still unsuitable for the target dataset.

Retraining-based methods. Due to the complex spatial structures in remote sensing images, CNN models that are pretrained or fine-tuned do not effectively reflect their unique property information. Therefore, researchers have started to train neural networks from scratch on raw remote sensing image datasets. Zhang et al. [30] proposed a gradient-boosted random convolutional network (GBRCN) framework for fusing neural networks, which introduced a deep integration framework for image scene classification. He et al. [31] presented an innovative hop-connected covariance (SCCov) network to solve the problem of remote sensing classification, which could achieve superior classification performance. This method required a mass of annotated samples, but the existing remote sensing image datasets were not large enough, which could cause overfitting problems.

In this study, we used the pretrained ResNet50 as the feature extractor and directly used the feature vector from the last fully connected layer of the network as the final representation of the image.

3. Methods

In this section, the framework of the proposed method is introduced, followed by a detailed description of the common space and label space in representation learning and, finally, the three loss functions used in this study, namely, discrimination loss in the common space, discrimination loss in the label space, and confusion loss.

(1) Framework of DHRL

As shown in Figure 2, the proposed framework adopts a duplex network to explore both the common space and label space to learn discriminative representations for RSIC. This method inputs a pair of images for training but only uses a single image for testing. The method consists of three main steps: First, in the feature extraction step, the semantic features of the images are extracted by the pretrained ResNet [20]. Then, the extracted features are input into the representation learning phase through several fully connected network layers to explore the consistency of the representations in the common space. Finally, relying on the assumption that common representations in the common space are optimal for classification, a linear classifier with a parameter matrix Q is used to predict the semantic categories of the input images in the label space, and the confusion score of the predicted results and the real labels is calculated using the confusion loss, which further optimizes the model.

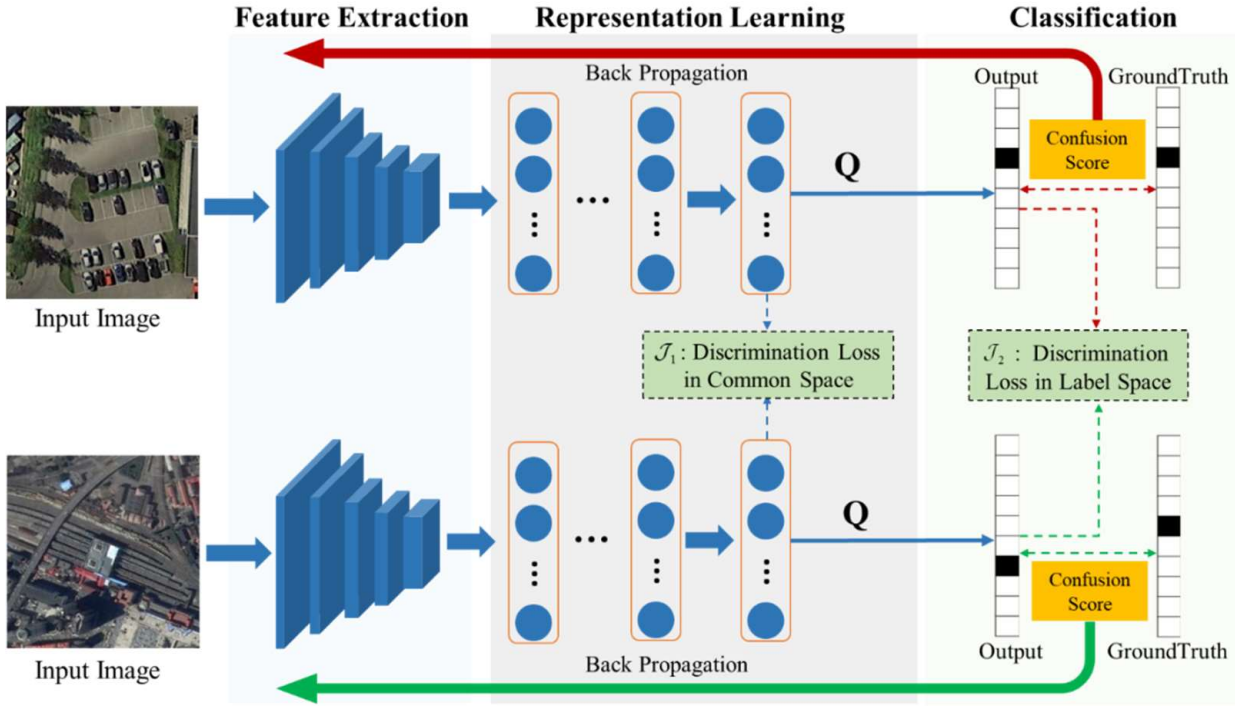

Figure 2. The framework of DHRL. First, the paired images are fed to a pretrained ResNet network for extracting the corresponding features. Second, the extracted features are further explored and mapped into a common representation space for reducing the intra-class scatter and enlarging the inter-class separation. Third, the obtained representation is used to predict the categories of the images and further facilitate the learning of discriminatory representations in the label space. Meanwhile, the confusion score is computed and added to the classification loss for guiding the discriminative representation learning via backpropagation.

Assume that there are n instances of image pairs, denoted as $\Psi = \left\{\left(\mathrm{x}_\mathrm{i}^{\alpha}, \mathrm{x}_\mathrm{i}^{\beta}\right)\right\}_{i=1}^{n}$, where $\mathrm{x}_\mathrm{i}^{\alpha}$ is the input image sample and $\mathrm{x}_\mathrm{i}^{\beta}$ is another image sample of the ith instance. $\mathrm{x}_\mathrm{i}^{\alpha}$ and $\mathrm{x}_\mathrm{i}^{\beta}$ are two randomly selected images from the dataset. If the image pairs have consistent labels, the features are constrained to be as similar as possible, but for image pairs with inconsistent labels, the features are constrained to be as different as possible. This enables the same category clusters to be more compact while allowing different category clusters to be as dispersed as possible. Each pair of instances $\left(\mathrm{x}_\mathrm{i}^{\alpha}, \mathrm{x}_\mathrm{i}^{\beta}\right)$ is assigned a semantic label vector $\left(y_\mathrm{i}^{\alpha}, y_\mathrm{i}^{\beta}\right)$, and $y_i = [y_{1i}, y_{2i}, \ldots, y_{ci}] \in R^c$, where c is the number of categories. If the ith instance belongs to the jth category, $y_{ij} = 1$; otherwise, $y_{ij} = 0$. Representation learning involves learning two functions for two inputs:

$$u_i = f(x_i^{\alpha}, Y_a) \in R^d, v_i = f\left(x_i^{\beta}, Y_\beta\right) \in R^d \tag{1}$$

where d is the dimensionality of the representation in the common representation space, u_i and v_i are the representations of instances x_i^{α} and x_i^{β} in the common space, and Y_a and Y_β are the trainable parameters of the two functions. This means that the similarity of samples from the same category is larger than the similarity of samples from different categories in the common space. In the following, the image representation matrix and the label matrix for all instances of Ψ are denoted as $U = [u_1, u_2, \ldots, u_n]$, $V = [v_1, v_2, \ldots, v_n]$, $Y^{\alpha} = [y_1^{\alpha}, y_2^{\alpha}, \ldots, y_3^{\alpha}]$, and $Y^{\beta} = [y_1^{\beta}, y_2^{\beta}, \ldots, y_3^{\beta}]$.

(2) Implementation Details

Deep neural networks are used to extract features directly from raw images, and various recent studies have shown that several excellent neural networks can be useful in RSIC tasks, such as AlexNet [32], VGG [33], and ResNet [20]. ResNet was proposed by He et al. in 2015 as a residual network. Usually, the deeper the network, the greater the amount of information that can be obtained and the richer the features. However, as the network gets deeper, it can cause problems such as gradient disappearance and gradient explosion. ResNet is an ultra-deep neural network, as it learns the residual representations between inputs and outputs, unlike the usual CNNs (AlexNet, VGG, etc.) that use participant layers to try to directly learn the mapping between inputs and outputs, and it greatly improves the accuracy; therefore, we chose ResNet as the backbone network for our proposed method.

ResNet consists of a diverse range of network layer depths. The most commonly encountered ones are 50 layers, 101 layers, and 152 layers; they are all constructed by stacking the previously mentioned residual modules together. In this study, the pretrained ResNet was used to extract the features of the conv5 layer from the original image, and features of $1 \times 1 \times 1204$ were finally obtained after pooling.

Due to the existence of intra-class diversity and inter-class similarity in RSIC, the main solution is to make instances of different classes as separate as possible while making instances of the same class as close as possible; therefore, it is necessary to measure the content similarity between different samples. Representation learning is an attempt to find a function that maps the obtained data samples into a common space where the similarity between them can be directly measured.

In the common space, samples from the same category should be similar, while samples from different categories should be dissimilar, as the similarity of samples from the same category is greater than the similarity of samples from different categories. Therefore, common representations need to be obtained, which are learned through several fully connected network layers after obtaining the features of images. By minimizing the discrimination loss in the common space, the intra-class distance is reduced while the inter-class distance is increased. The discrimination loss of all samples from both images in the common space is measured directly:

$$\begin{aligned} \mathcal{J}_1 = & \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}\left(log\left(1+e^{\Lambda_{ij}}\right) - P_{\alpha\beta}\Lambda_{ij}\right) \\ & + \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}\left(log\left(1+e^{\mathrm{T}_{ij}}\right) - P_{\alpha\alpha}\mathrm{T}_{ij}\right) \\ & + \tfrac{1}{n^2}\textstyle\sum_{i,j=1}^{n}\left(log\left(1+e^{\mathrm{Y}_{ij}}\right) - P_{\beta\beta}\Phi_{ij}\right) \end{aligned} \tag{2}$$

where $\Lambda_{ij} = \frac{1}{2}cos\left(u_i, v_j\right)$, $\mathrm{T}_{ij} = \frac{1}{2}cos\left(u_i, \mu_j\right)$, $\Phi_{\mathrm{ij}} = \frac{1}{2}\cos\left(\nu_{\mathrm{i}}, \nu_{\mathrm{j}}\right)$, $cos(\cdot)$ is a cosine function used to compute the similarity of two input vectors, $\mathrm{P}_{\alpha\beta} = \mathrm{L}\left\{\mathrm{u_i}, \mathrm{v_j}\right\}$, $P_{\alpha\alpha} = L\left\{u_i, \mu_j\right\}$, $P_{\beta\beta} = L\left\{\nu_i, v_j\right\}$, and $L(\cdot)$ is 1 when the two samples are representations of intra-class samples and 0 otherwise. Each term of Equation (2) is the negative log-likelihood of the sample similarities, and minimizing the negative log-likelihood is equivalent to maximizing the likelihood. It can be seen that the larger the similarities $\cos(\cdot)$ are, the larger $p(1 \mid u, v)$ will be, which means that the image samples should be classified as similar, and vice versa. The first term of Equation (2) measures the similarities between the two image samples, and the second and third terms measure the similarities between samples of the interior of the image. Therefore, Equation (2) is a reasonable measure of similarity for common representations and is an effective criterion for learning discriminative features.

The label space is mainly used to classify the obtained representations. Due to the supervised methods, label information is used to distinguish samples from different semantic categories in order to learn more differentiated generic representations. Figure 3 demonstrates a simple example for the representations in three spaces. To preserve the discrimination of samples from different categories after the feature projection, it is assumed that the common representations are ideal for classification, and a linear classifier with a parameter matrix Q is used to predict the labels of the samples projected in the common space. This classifier takes the representations of the training data in the common

space and generates a predicted label of a c-dimensional vector for each sample. For the image sample x_i, the input of linear classifier Q is the learned feature representation u_i in the common space, and the output of linear classifier Q is the predicted label y_i. In this study, the following objective function was used to measure the discriminative loss in the label space:

$$\mathcal{J}_2 = \frac{1}{n}\left|\left|Q^T U - Y^{\alpha}\right|\right|_F + \frac{1}{n}\left|\left|Q^T V - Y^{\beta}\right|\right|_F \tag{3}$$

where $||\cdot||_F$ is the Frobenius norm, and Q denotes the projection matrix of the linear classifier.

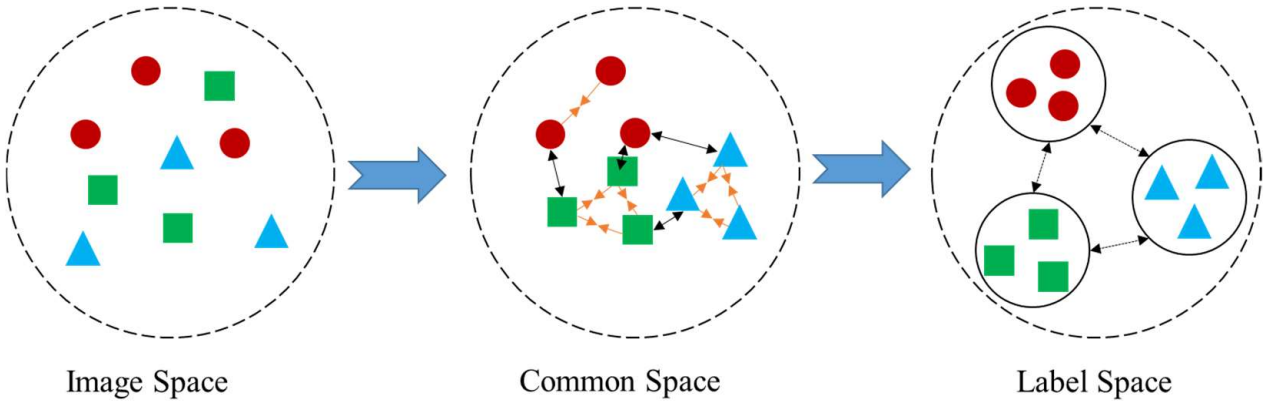

Figure 3. Samples in dual-level spaces. In the common space, samples with similar features are brought together and samples with different features are separated, thereby increasing the inter-category distance and decreasing the intra-category distance.

To better allow the ground truth to be supervised, confusion loss is proposed in this study. The purpose of the confusion loss is to ensure that the predicted labels are closer to the true labels, especially for samples that can easily be predicted incorrectly. By calculating the confusion between them via backpropagation to the loss function, the loss can adjust the model so that it gives more attention to the samples that can easily be predicted incorrectly, minimizing the confusion loss and learning more discriminative features via backpropagation to achieve more discriminative representations. The following gives the confusion loss:

$$L_c = \left(1 - \sum_{k=1}^{n} \frac{e^{{y_k}^T y_k^*}}{\sum_{l=1}^{n} e^{{y_l}^T y_l^*}}\right) \times \left(-\sum_{k=1}^{n} {y_k}^T log\left(\frac{e^{y_k^*}}{\sum_{l=1}^{n} e^{y_l^*}}\right)\right) \tag{4}$$

where n is the number of categories, y_i^* is the final output, and y_i is the ground truth. Combining Equations (2)–(4), the objective function of the proposed method is as follows:

$$\mathcal{J} = \mathcal{J}_1 + \mathcal{J}_2 + L_c \tag{5}$$

Algorithm 1 provides an overview of DHRL. The objective function of DHRL in Equation (5) can be optimized by using a stochastic gradient descent optimization algorithm.

Algorithm 1. Framework of the proposed DHRL method.

Input: The training dataset $\Psi = \left\{ \left(x_i^{\alpha}, x_i^{\beta} \right) \right\}_{i=1}^{n}$ the label matrix Y, the dimensionality of common space d, the batch size n_b, the learning rate τ, and the maximal number of epochs $\aleph$.
Output: Predict the label for the input image.

1. Randomly initialize the parameters of the two subnetworks Y_a and Y_β, and the parameters of the linear classifier Q.
2. For $t = 1, 2, ..., \aleph$, do
3. For $l = 1, 2, \ldots, [\frac{n}{n_b}]$, do
4. Randomly sample n_b image pairs samples from Ψ to construct a mini-batch.
5. Compute the representations u_i and v_i for the samples in the mini-batch through forward propagation.
6. Calculate the result of the objective function in Equation (5).
7. Update the parameters of the subnetworks Y_a and Y_β and the linear classifier by minimizing the loss function.
8. End for
9. End for
10. Calculate the network output according to $Y^{\alpha *} = Q^T U$, $Y^{\beta *} = Q^T V$

4. Results

4.1. Datasets

NWPU-RESISC45 dataset [20]: Our proposed method was trained on the NWPU-RESISC45 dataset. This dataset was generated by the Northwestern Polytechnical University research team in 2017 and includes 31,500 remote sensing images and 45 scene classes. Each scene class consists of 700 images, where the dimensions of each image are 256 × 256. The spatial resolution of the majority of the images is 30~0.2 m/pixel, and some images of specific terrains may have a lower resolution, such as lakes, islands, and regular mountains. The dataset encompasses a diverse array of scene categories, with each category preserving substantial internal diversity while also displaying similarities to other categories.

AID dataset [34]: This dataset was created by Wuhan University in 2017. It has an image size of 600 × 600 and consists of 30 scene categories. Each category contains a maximum of 400 images and a minimum of 200 images, and the image resolution ranges from 0.5 to 8 m.

4.2. Implementation Details

In this work, our experimental verifications were carried out on a computer with a GTX TITAN GPU with 12G, and the algorithms were implemented with TensorFlow [35] and Keras [36]. The details of the parameter settings for the proposed DHRL were as follows. There were 30 epochs for the model, and the Adam [37] optimizer was applied with 10^{-5}.

To assess the merits of our proposed algorithm in comparison with other state-of-the-art algorithms, it was imperative to ensure uniformity in data segmentation across all compared and benchmark methods. Therefore, using different training ratios for the different datasets allowed for a better analysis of the strengths and weaknesses of the method proposed in this study. For the NWPU-RESISC45 dataset, the training ratios were set to 10% and 20%, with the remaining 90% and 80% being used for testing. The training images were randomly rotated by 30° and flipped both horizontally and vertically. For the AID dataset, the training ratios were set to 20% and 50%, with the remaining 80% and 50% being used for testing. The training ratio represented the proportion of images in the dataset used for training. For example, 10% meant that 10% of the images were used for training. For a fair comparison, we used the same training ratios as those of the state-of-the-art methods for the NWPU-RESISC45 dataset and the AID dataset.

4.3. Comparison with Other State-of-the-Art Methods

4.3.1. NWPU-RESISC45 Dataset

For the NWPU-RESISC45 dataset, the method proposed in this study was compared with existing methods (CNN-CapsNet [38], SCCov [31], ADFF, Siamese ResNet50 [39], FDPResnet [40], DDRL-AM [41], SF-CNN [18], and HABFNet [42]). The results are presented in Table 1. The experimental results showed that the test accuracy of our proposed method was 96.03 and 96.32 when the training ratio was 10% and 20%, respectively. CNN-CapsNet took full advantage of both CNN and CapsNet models, and its overall accuracy was 7% and 3.72% lower than that of the proposed DHRL method at the 10% and 20% training ratios. SCCov had accuracies of 89.30 and 92.10 at training ratios of 10% and 20%, which were 6.73% and 4.22% lower than those of DHRL. ADFF employed an attention mechanism, and its accuracy was 5.45% and 4.41% lower than that of DHRL. Siamese ResNet50 combined CNN identification and validation models, and it reduced the accuracy by 4.04% at a training ratio of 20%. FDPResnet is a fusion of the DCNN and the new and effective Extensive Learning System (BLS) for fast depth perception networks. It reduced the accuracy by 3.71% and 0.92% compared to DHRL. DDRL-AM [41] implemented depth-differentiated representation learning based on attention maps, and its accuracy was 3.86% lower than that of the method proposed in this study. HABFNet was a variety of methods based on the feature fusion framework of hierarchical attention and bilinear pooling. The proposed method improved upon its accuracy by 3.28% and 1.78%. Thus, the performance of DHRL on the NWPU-RESISC45 dataset was more effective than the performance of the current advanced methods, which confirmed the effectiveness of DHRL.

Table 1. Classification accuracy of different methods on the NWPU-RESISC45 dataset.

Methods	Training Ratio (%)	
	10%	**20%**
CNN-CapsNet [38]	89.03	92.6
SCCov [31]	89.30	92.10
ADFF	90.58	91.91
Siamese ResNet50 [39]	—	92.28
FDPResnet [40]	92.32	95.40
DDRL-AM [41]	92.17	92.46
SF-CNN [18]	89.89	92.55
HABFNet [42]	92.75	94.54
RCOVBOVW [43]	90.25	93.27
HFFCNN [44]	87.01	90.14
MGSNet [45]	92.4	94.57
DHRL (ours)	**96.03**	**96.32**

To enhance the comprehension of the performance exhibited by DHRL, a confusion matrix was constructed to visually depict the accuracy of the classification. The result is shown in Figure 4. Each row in the matrix corresponds to the actual category, whereas each column pertains to the predicted category. Cells along the diagonal signify accurate predictions, while off-diagonal cells signify errors. The color within each cell denotes the cumulative count and percentage of prediction instances, with correct categorizations being arranged sequentially along the diagonal axis from left to right.

As shown in Figure 4, when the training ratio of the dataset was 20%, a classification accuracy of over 90% was achieved for most of the categories, which further indicated the effectiveness of DHRL. The classification accuracy was also obtained for categories with intra-category diversity, such as train stations and churches. For categories with inter-category similarity, such as churches and palaces, airports and lawns, and tennis and basketball courts, the classification accuracy was lower than that for other categories, but the error rate was low, indicating that DHRL was able to effectively solve the problem of intra-category diversity and inter-category similarity.

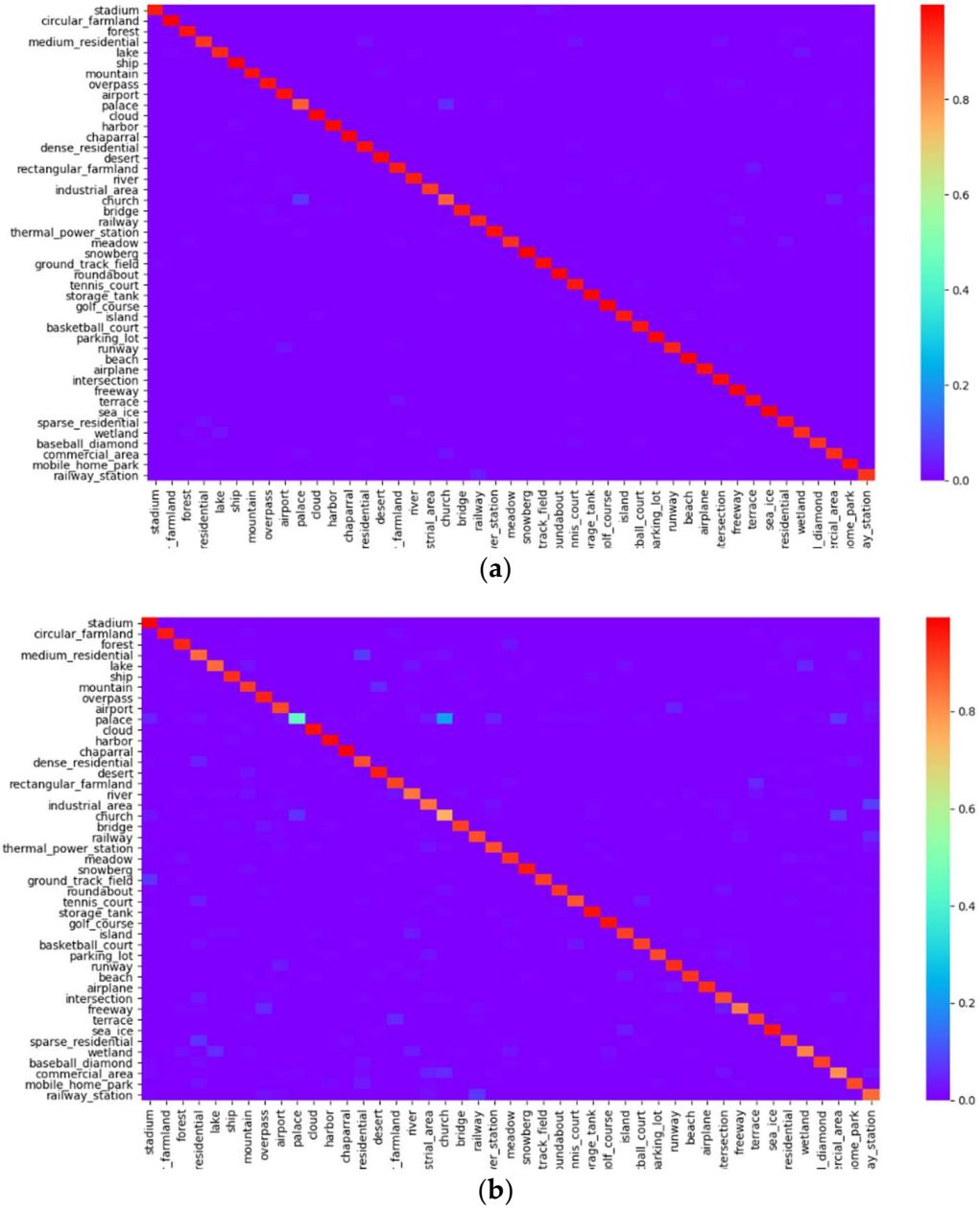

Figure 4. Confusion matrix of the proposed DHRL method on the NWPU-RESISC45 dataset. (**a**) Training ratio of 20%. (**b**) Training ratio of 10%.

4.3.2. AID Dataset

As shown in Table 2, the proposed DHRL method also yielded excellent results on the AID dataset. At 20% and 50% training ratios, the accuracy of DHRL was 93.08% and 96.54%, respectively. When the training ratio was 20%, DHRL outperformed most methods, improving the accuracy by 0.72%, 0.88%, and 2.83% compared to DDRL-AM [41], GBNet [46], and VGG_VD16+SAFF [47], and it had a similar accuracy to that of SCCov. ADFF represents an attention-based deep feature fusion framework comprising three key components: attentional mapping guided by gradient-weighted class activation mapping (GradCAM), multiplicative fusion of deep features, and the utilization of a center-based cross-entropy loss function. It was superior to our proposed DHRL method. However, at a 50% training ratio, DHRL outperformed all existing comparison methods, improving the accuracy by 0.44%, 1.09%, 1.79%, 4.74%, 1.06%, and 2.71% compared to SCCov [31], FACNN [27], ADFF, MCNN [42], GBNet [46], and VGG_VD16+SAFF [47]. When there was

a small amount of training data, the multiple feature fusion strategy of ADFF was able to obtain more information for classification and achieve a slightly higher performance. When there was a large amount of training data, the confusion loss optimization strategy of the DHRL method was able to learn more discriminative representations for the samples that could be easily misclassified; therefore, the performance of DHRL was the best at a training ratio of 50%. Thus, the advantage of DHRL was more pronounced in the case of a larger amount of training data.

Table 2. Classification accuracy of different methods on the AID dataset.

Methods	Training Ratio (%)	
	20%	**50%**
SCCov [31]	93.12	96.10
FACNN [27]	—	95.45
ADFF	**93.68**	94.75
DDRL-AM [41]	92.36	—
MCNN [42]	—	91.80
GBNet [46]	92.20	95.48
VGG_VD16+SAFF [47]	90.25	93.83
HFFCNN [44]	93.08	95.32
DHRL (ours)	93.08	**96.54**

4.4. Ablation Study

To investigate the effectiveness of the proposed method, we designed experiments to evaluate the performance of the three losses ($\mathcal{J}_1$, $\mathcal{J}_2$, and L_c). The loss function of the DHRL method consisted of three components, which were used to characterize the discrimination loss in the common space, the discrimination loss in the label space, and the confusion loss of the classification. We developed three variations of the objective function of the proposed DHRL method: DHRL1 without $\mathcal{J}_1$, DHRL2 without $\mathcal{J}_2$, and DHRL3 without L_c. We evaluated the performance of these variations on the NWPU-RESISC45 dataset at training ratios of 10% and 20%. Table 3 shows the experimental classification results.

Table 3. Classification accuracy of the DHRL method and its three variations on the NWPU-RESISC45 dataset.

Methods	Training Ratio (%)	
	10%	**20%**
DHRL1	93.22	94.53
DHRL2	82.19	85.67
DHRL3	91.35	92.82
Full DHRL	**96.03**	**96.32**

The result in Table 3 demonstrates that the performance of the full DHRL method was the best, which indicated that all three components of the loss function contributed to the scene image classification accuracy. We also observed that the reduction in the classification accuracy of DHRL2 was the largest because the second component $\mathcal{J}_2$ optimized the discrimination loss directly in the label space. Based on the experimental results, it was proven that optimizing both the discrimination loss and confusion loss in the objective function was an effective learning method.

5. Conclusions

(1) In this study, we proposed a dual hierarchical representation learning approach for the problem of large dissimilarities in the same class and small dissimilarities between different classes in remote sensing image classification with the aim of exploring the dual hierarchical space, including a common space and a label space, to learn

differentiated representations for remote sensing image classification by assessing the distinctions between different classes within the common space.

(2) Experimental evaluations conducted on challenging datasets demonstrated that the outcomes of the method proposed in this study surpassed those of existing state-of-the-art methods. This substantiated the efficacy and validity of the proposed approach.

(3) In our future work, we will extend the proposed method to multi-source remote sensing image classification. Fusing multi-source image features in the common space is a valuable strategy.

Author Contributions: Conceptualization, X.Y. and J.Z.; methodology, X.Y. and H.L.; software, S.P., W.W. and X.L.; validation, S.P., W.W. and X.L.; writing—original draft preparation, X.Y.; writing—review and editing, H.L.; supervision, J.Z.; funding acquisition, X.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Basic Research Program of Shaanxi (Program No. 2023-YBGY-028).

Data Availability Statement: Two datasets were used in this work: The NWPU-RESISC45 dataset is openly available at https://gcheng-nwpu.github.io/#Datasets accessed on 1 May 2022, and the AID dataset is openly available at https://captain-whu.github.io/AID/ accessed on 1 May 2022.

Acknowledgments: We would like to thank the handling editors and the anonymous reviewers for their careful and helpful comments and suggestions.

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

References

1. Cheng, G.; Guo, L.; Zhao, T.; Han, J.; Li, H.; Fang, J. Automatic landslide detection from remote-sensing imagery using a scene classification method based on BoVW and pLSA. *Int. J. Remote Sens.* **2013**, *34*, 45–59. [CrossRef]
2. Fan, J.; Chen, T.; Lu, S. Unsupervised feature learning for land-use scene recognition. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2250–2261. [CrossRef]
3. Longbotham, N.; Chaapel, C.; Bleiler, L.; Padwick, C.; Emery, W.J.; Pacifici, F. Very high resolution multiangle urban classification analysis. *IEEE Trans. Geosci. Remote Sens.* **2011**, *50*, 1155–1170. [CrossRef]
4. Swain, M.J.; Ballard, D.H. Color indexing. *Int. J. Comput. Vis.* **1991**, *7*, 11–32. [CrossRef]
5. Clausi, D.A.; Deng, H. Design-based texture feature fusion using Gabor filters and co-occurrence probabilities. *IEEE Trans. Image Process.* **2005**, *14*, 925–936. [CrossRef] [PubMed]
6. Zhang, L.; Huang, X.; Huang, B.; Li, P. A pixel shape index coupled with spectral information for classification of high spatial resolution remotely sensed imagery. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 2950–2961. [CrossRef]
7. Haralick, R.M.; Shanmugam, K.; Dinstein, I.H. Textural features for image classification. *IEEE Trans. Syst. Man Cybern.* **1973**, *3*, 610–621. [CrossRef]
8. Oliva, A.; Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* **2001**, *42*, 145–175. [CrossRef]
9. Lowe, D.G. Distinctive image features from scale-invariant key points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
10. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 886–893. [CrossRef]
11. Perronnin, F.; Sanchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In Proceedings of the European Conference on Computer Vision, Heraklion, Greece, 5–11 September 2010; pp. 143–156.
12. Jegou, H.; Perronnin, F.; Douze, M.; Sánchez, J.; Pérez, P.; Schmid, C. Aggregating local image descriptors into compact codes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *34*, 1704–1716. [CrossRef] [PubMed]
13. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **2006**, *2*, 2169–2178.
14. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
15. Penatti, O.A.B.; Nogueira, K.; Dos Santos, J.A. Do deep features generalize from everyday objects to remote sensing and aerial scenes domains? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015; pp. 44–51.
16. Maggiori, E.; Tarabalka, Y.; Charpiat, G.; Alliez, P. Convolutional Neural networks for large-scale remote-sensing image classification. *IEEE Trans. Geosci. Remote Sens.* **2016**, *55*, 645–657. [CrossRef]

17. Liu, Y.; Zhong, Y.; Qin, Q. Scene classification based on multiscale convolutional neural network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 7109–7121. [CrossRef]
18. Xie, J.; He, N.; Fang, L.; Plaza, A. Scale-free convolutional neural network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6916–6928. [CrossRef]
19. Castelluccio, M.; Poggi, G.; Sansone, C.; Verdoliva, L. Land use classification in remote sensing images by convolutional neural networks. *arXiv* **2015**, arXiv:1508.00092.
20. Cheng, G.; Han, J.; Lu, X. Remote sensing image classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]
21. 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, 26 June–1 July 2016; pp. 770–778.
22. Rawat, W.; Wang, Z. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Comput.* **2017**, *29*, 2352–2449. [CrossRef] [PubMed]
23. Agrawal, P.; Girshick, R.; Malik, J. Analyzing the performance of multilayer neural networks for object recognition. In Proceedings of the Computer Vision—ECCV 2014, Zurich, Switzerland, 6–12 September 2014; pp. 329–344. [CrossRef]
24. Guo, Y.; Liu, Y.; Georgiou, T.; Lew, M.S. A review of semantic segmentation using deep neural networks. *Int. J. Multimed. Inf. Retr.* **2018**, *7*, 87–93. [CrossRef]
25. Hu, F.; Xia, G.-S.; Hu, J.; Zhang, L. Transferring Deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery. *Remote Sens.* **2015**, *7*, 14680–14707. [CrossRef]
26. Cheng, G.; Li, Z.; Yao, X.; Guo, L.; Wei, Z. Remote sensing image scene classification using bag of convolutional features. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 1735–1739. [CrossRef]
27. Lu, X.; Sun, H.; Zheng, X. A feature aggregation convolutional neural network for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 7894–7906. [CrossRef]
28. Liu, Y.; Suen, C.Y.; Liu, Y.; Ding, L. Scene classification using hierarchical Wasserstein CNN. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 2494–2509. [CrossRef]
29. Wang, Q.; Liu, S.; Chanussot, J.; Li, X. Scene classification with recurrent attention of VHR remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2018**, *57*, 1155–1167. [CrossRef]
30. Zhang, F.; Du, B.; Zhang, L. Scene classification via a gradient boosting random convolutional network framework. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 1793–1802. [CrossRef]
31. He, N.; Fang, L.; Li, S.; Plaza, J.; Plaza, A. Skip-connected covariance network for remote sensing scene classification. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *31*, 1461–1474. [CrossRef]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **2012**, *25*, 1097–1105. [CrossRef]
33. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
34. Xia, G.S.; 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]
35. Abadi, M.; Barham, P.; Chen, J.; Davis, J.; Dean, J.; Devin, M.; Ghemawat, S.; Irving, G.; Isard, M.; Kudlur, M.; et al. TensorFlow: A system for Large-Scale machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
36. Ketkar, N. Introduction to keras. In *Deep Learning with Python: A Hands-On Introduction*; Apress: Berkeley, CA, USA, 2017; pp. 97–111.
37. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using CNN-CapsNet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
39. Liu, X.; Zhou, Y.; Zhao, J.; Yao, R.; Liu, B.; Zheng, Y. Siamese convolutional neural networks for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2019**, *16*, 1200–1204. [CrossRef]
40. Dong, R.; Xu, D.; Jiao, L.; Zhao, J.; An, J. A Fast Deep Perception Network for Remote Sensing Scene Classification. *Remote Sens.* **2020**, *12*, 729. [CrossRef]
41. Li, J.; Lin, D.; Wang, Y.; Xu, G.; Zhang, Y.; Ding, C.; Zhou, Y. Deep discriminative representation learning with attention map for scene classification. *Remote Sens.* **2020**, *12*, 1366. [CrossRef]
42. Yu, D.; Guo, H.; Xu, Q.; Lu, J.; Zhao, C.; Lin, Y. Hierarchical Attention and Bilinear Fusion for Remote Sensing Image Scene Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6372–6383. [CrossRef]
43. Chen, X.; Zhu, G.; Liu, M. Bag-of-visual-words scene classifier for remote sensing image based on region covariance. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]
44. Xu, K.; Deng, P.; Huang, H. Mining Hierarchical Information of CNNs for Scene Classification of VHR Remote Sensing Images. *IEEE Trans. Big Data* **2022**, *9*, 542–554. [CrossRef]
45. Wang, J.; Li, W.; Zhang, M.; Tao, R.; Chanussot, J. Remote sensing scene classification via multi-stage self-guided separation network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5615312.

46. Sun, H.; Li, S.; Zheng, X.; Lu, X. Remote sensing scene classification by gated bidirectional network. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 82–96. [CrossRef]
47. Cao, R.; Fang, L.; Lu, T.; He, N. Self-Attention-Based Deep Feature Fusion for Remote Sensing Scene Classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 43–47. [CrossRef]