electronics

Article

FCIHMRT: Feature Cross-Layer Interaction Hybrid Method Based on Res2Net and Transformer for Remote Sensing Scene Classification

Yan Huo [1,2,3,†], Shuang Gang [1,2,3,†] and Chao Guan [1,2,3,*]

[1] Institute of Carbon Neutrality Technology and Policy, Shenyang University, Shenyang 110044, China
[2] Northeast Geological S&T Innovation Center of China Geological Survey, Shenyang 110034, China
[3] Key Laboratory of Black Soil Evolution and Ecological Effect, Ministry of Natural Resources, Shenyang 110034, China
[*] Correspondence: gc471603869@syu.edu.cn
[†] These authors contributed equally to this work.

Abstract: Scene classification is one of the areas of remote sensing image processing that is gaining much attention. Aiming to solve the problem of the limited precision of optical scene classification caused by complex spatial patterns, a high similarity between classes, and a high diversity of classes, a feature cross-layer interaction hybrid algorithm for optical remote sensing scene classification is proposed in this paper. Firstly, a number of features are extracted from two branches, a vision transformer branch and a Res2Net branch, to strengthen the feature extraction capability of the strategy. A novel interactive attention technique is proposed, with the goal of focusing on the strong correlation between the two-branch features, to fully use the complementing advantages of the feature information. The retrieved feature data are further refined and merged. The combined characteristics are then employed for classification. The experiments were conducted by using three open-source remote sensing datasets to validate the feasibility of the proposed method, which performed better in scene classification tasks than other methods.

Keywords: vision transformer; remote sensing image; Res2Net; scene classification

Citation: Huo, Y.; Gang, S.; Guan, C. FCIHMRT: Feature Cross-Layer Interaction Hybrid Method Based on Res2Net and Transformer for Remote Sensing Scene Classification. *Electronics* **2023**, *12*, 4362. https://doi.org/10.3390/electronics12204362

Academic Editor: Hüseyin Kusetogullari

Received: 13 September 2023
Revised: 17 October 2023
Accepted: 18 October 2023
Published: 20 October 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The rapid development of remote sensing technologies with satellites and unmanned aerial vehicles has produced a large number of high-resolution remote sensing images with rich scenes. These images can be applied in forest state assessment, urban planning, ecological environment monitoring, and many other applications [1]. By extracting and evaluating the properties of the remote sensing images, the primary application goal is to reliably identify the target categories included in the images, such as buildings, forests, and wetlands. A significant area of remote sensing research is scene classification [2]. Effective scene differentiation is especially crucial since remote sensing images have substantial intra-class differences and small inter-class differences.

In general, the scene classification approaches for remote sensing images can be divided into three primary kinds: low-level methods, mid-level methods, and high-level feature methods [3]. For early traditional scene classification methods, low and mid-level features are mostly obtained from remote sensing images manually. Low-level methods usually use handcrafted features, e.g., local binary patterns, scale-invariant feature transform features, and histograms of oriented gradients [4]. Handcrafted features perform well in remote sensing images with a neat texture and uniform spatial distribution, but they have difficulties in depicting the semantic information of complicated images. Mid-level methods use statistical calculations or the coding of low-level features, e.g., bag of visual words [5], and probabilistic topic models [3]. However, low-level and mid-level methods need a large amount of experience and time. These manual methods have the shortcomings of providing little information and having a low effectiveness, and they have difficulties in satisfying the needs of remote sensing applications. Therefore, high-level methods using deep learning are applied to efficiently extract visual information for scene classification.

Recently, it was found that convolutional neural networks (CNNs) can significantly improve the efficiency of remote sensing scene classification [6,7]. Different from traditional classification methods, CNNs (e.g., AlexNet, VGGNet, GoogLeNet, ResNet, and DenseNet) [8] have the characteristics of local perception and weight sharing. Only using the features of the CNN middle layer or the full connection layer as the image feature representation will ignore the complementary advantages of different levels of information, resulting in a low portrayal capability of the network. Although algorithms based on CNNs capture local texture, they are not able to depict the global structure of scene images, so the classification accuracy will encounter a bottleneck. Bahdanau et al. [9] designed a new attention method, which aimed to solve the problem of a too-long training time. This model not only solves the information bottleneck in machine translation, but also alleviates the gradient disappearance of recurrent neural networks (RNNs) in long-distance dependency. Recent advancements in the field of natural language processing (NLP) have highlighted the potential of a technique with a self-attention mechanism, named Transformer [10], which has the ability to refresh the parameters of a deep learning model using global computing on the input sequence. Inspired by the Transformer application in NLP, Dosovitskiy et al. [11] applied an attention mechanism with a vision transformer (ViT) for image classification and recognition tasks. This model can effectively extract the long-distance dependent (image structure) information of natural images and overcome the conduction bias in CNNs. To enhance the recognition rate of scene features, Deng et al. [12] proposed an efficient combined approach by integrating ViT into a CNN model.

To obtain the overall structure and local texture details of scene images at the same time, a feature cross-layer interaction hybrid method based on Res2Net and Transformer, called FCIHMRT, is proposed in this paper. According to the characteristics of spatial information diversity, small objects and inter-class diversity of remote sensing images, an effective attention mechanism is introduced into the designed model. Meanwhile, multiple features are fused for classification to enhance feature utilization. The main contributions of this paper are summarized as follows:

(1) A novel hybrid network is developed to effectively combine the advantages of transformer and Res2Net in order to extract numerous features with high-value information and to increase classification efficiency.

(2) A cross-layer interactive module is proposed to integrate multi-feature information. This module can fuse the two extraction features of the ViT branch and the Res2Net branch to enhance the representation ability.

(3) A new interactive attention mechanism is designed for focusing on the deep correlation between the two-path features. The mechanism uses two global pooling operations to reduce the dimension of the channel. Attention weights are used to enhance the feature response of valuable feature information.

(4) Training on the three public datasets UC-Merced (UCM), Aerial Image Dataset (AID), and NWPU-RESISC45 (NWPU) is completed. The results of the experiments show that the proposed approach outperforms the current advanced CNN techniques.

The remainder of this article is structured as follows. The relevant work is introduced in Section 2, and the designed algorithm is covered in Section 3. Section 4 illustrates the experimental investigation and relevant comparisons. In Section 5, the conclusions are discussed.

2. Related Work

2.1. CNN-Based Methods

A convolutional neural network (CNN) [13] is a special artificial neural network structure, which has a wide range of applications in image recognition, speech recognition, natural language processing, etc. The characteristic of CNNs is that they can automatically extract the features of input data, so as to realize the efficient classification and recognition of the input data. Generally, CNNs include a convolutional layer, pooling layer, fully connected layer, and activation function. The convolutional layer extracts features from input data through the convolutional operation, which is a mathematical operation that generates a new feature map by sliding a convolution kernel over the input data and calculating the dot product of the convolution kernel with a local region of the input data. The pooling layer is used to reduce the dimensions of the feature map, thus reducing the amount of computation. The fully connected layer integrates the features extracted by the convolutional layer and the pooling layer, and it performs nonlinear transformations through the activation function to output the classification result. The activation function is used to introduce nonlinear factors so that the neural network can fit complex nonlinear relationships.

CNN-based methods have been the predominant technique used in scene classification due to their remarkable performance [14]. The CNN-based methods includes three groups: improved existing CNNs, transfer learning, and generative adversarial network [15], which are listed in Table 1 along with the general benefits and limitations of each group. The first group comprises improved CNN methods. Lu et al. [16] suggested a complete supervised feature-encoding technique using a CNN to incorporate feature learning and aggregation for examining semantic labeling data. To minimize the high-dimensional characteristics, Li et al. [17] presented an enhanced bilinear pooling technique based on a compact bilinear CNN framework. CNNs are also coupled with other networks to increase the precision of scene classification. For the purpose of classifying remote sensing scenes, Zhang et al. [18] established a hybrid feature learning strategy using a CNN and capsule networks. Peng et al. [19] presented a multi-output model for classifying scenes using a graph neural network and CNN with a combined loss using the backpropagation process. Although the improved CNN methods can perform classification tasks well, it is difficult to improve their performance further due to the excessive dependence on local spatial information [20]. The second group comprises transfer learning methods. For instance, Wang et al. [21] developed two promising architectures for collecting generic features from pre-trained CNNs for scene categorization. Additionally, it offers a baseline for adapting pre-trained CNNs for various remote sensing applications. In order to address input feature disparities between the target and source datasets, Zhao et al. [22] presented a heterogeneous methodology for using transferring CNNs in remote-sensing scene classification tasks. The pre-trained CNN network is utilized as a feature extractor on the chosen target dataset. Wang et al. [23] developed an adaptive learning technique for transferring CNNs to determine which important information should be transferred to the scene categorization model. The third group of scene classification uses generative adversarial networks (GANs) [24]. Han et al. [25] presented a scene categorization methodology for producing high-resolution annotated samples using remote sensing images, and it is based on GAN. In order to provide remote sensing image samples with label information, Ma et al. [26] constructed a supervised progressive evolving conditional GAN.

Table 1. CNN-based methods for remote sensing image classification.

Group	Benefit	Limitation
Improved CNN methods	Flexible to model	Excessive dependence on local spatial information
Transfer learning methods	Reduce the cost of training	Not easy to converge model parameters
Generative adversarial network methods	Produce high quality images	Difficult to achieve training balance

2.2. Attention-Based Methods

Attention-based techniques for remote sensing scene classification of high-level features are effectively used to learn deeper feature information [27,28], including conventional attention-based and transformer-based methods. Conventional attention-based methods are usually applied. To concentrate on the useful features, Wang et al. [29], for instance, built a deformable CNN based on the spatial and channel attention processes. Tian et al. [30] created a multiscale dense convolutional model with a squeeze-and-excitation attention mechanism to successfully establish the association between features and enrich the feature channels with helpful information. To obtain more precise feature information, Wang et al. [31] suggested using deep convolution neural networks that utilized a channel-wise attention method. Shen et al. [32] used dual CNNs to extract features and introduced a spatial attention mechanism into the classification model to avoid irrelevant information. Transformer-based approaches, which resort to a self-attention mechanism to determine the link between sequence parts, fall under the second category. With the purpose of enhancing the quality of the feature expression, Yu et al. [33] developed an innovative hybrid capsule-based ViT model that takes advantage of three feature semantics. To enhance the model learning ability, Zhang et al. [34] constructed an innovative remote sensing image classification approach based on a CNN and transformer with a multi-head self-attention layer. Sha et al. [35] created a multi-instance ViT framework for scene recognition by fusing multiple instance learning with ViT to take advantage of the important local features. Recently, Wang et al. [36] suggested a new architecture of plug-and-play CNN features integrated with a hybrid vision transformer by combining the benefits of CNN and ViT.

3. The Proposed Method

3.1. Framework Overview

The presented model of scene classification (FCIHMRT) comprises a processing branch based on ViT (left side) and a processing branch based on Res2Net (right side) for feature extraction, as depicted in Figure 1. In the branch based on ViT, the image is first divided into a number of image blocks, and their corresponding spatial positions are embedded. ViT encoders are used to process the embedded image blocks. The last feature is output by the ViT layers with the concat and reshape operations. In the other branch, the processing branch based on Res2Net is first utilized. The four processing blocks with Res2Net are utilized to extract features from scene images. Thus, ViT and Res2Net features are, respectively, processed with the convolution and pooling layer in the proposed cross-layer interaction module. They are also fused into a feature map in the attention module, which is used to produce the feature representation weights for dual-path features. In addition, the two-path features are concatenated with the extracted features of the other path in the cross-layer interaction module before weighting. Finally, the fused feature is fed to the Softmax classifier for classification. The prediction label of the scene image is output.

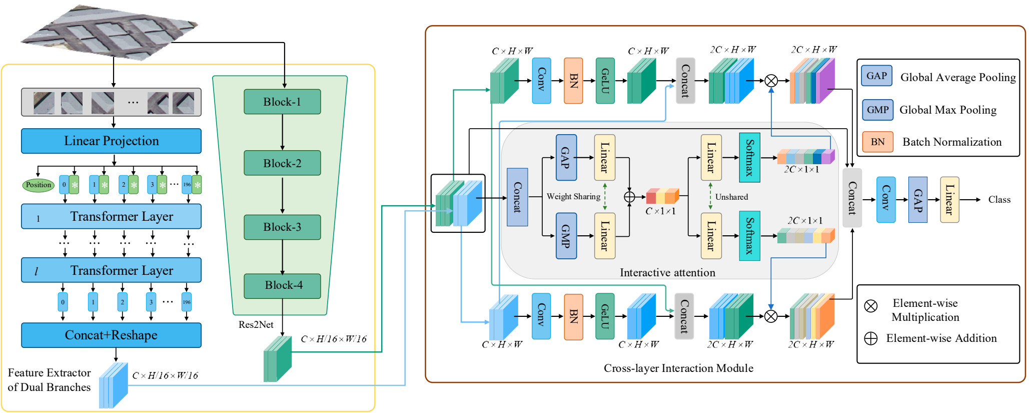

Figure 1. Overall structure of our proposed method.

3.2. Feature Extraction of Dual Branches

A structure with dual branches is proposed in order to address the issue of one CNN being incapable of feature extraction. Here, we build ViT and Res2Net as two efficient feature extractors.

3.2.1. ViT Branch

A ViT encoder consists of one stack of the same layer, each layer is composed of two sub-layers, and a multi-head self-attention (MSA) and multi-layer perceptron (MLP) modules, as displayed in Figure 2. Before the input data goes into each sublayer, layer normalization (LN) is used for normalization processing. After each sublayer, the obtained data are fused directly with the inputs using the residual connection. Finally, after L layers of network coding, the first element of the sequence x_L^0 is sent into the category header composed of the MLP so as to predict the category y of the image. The intermediate variable x_l' and the output x_l of the l-th layer are expressed as

$$x_l' = \text{MSA}(\text{LN}(x_{l-1})) + x_{l-1},\ l = 1, 2, \ldots, L \tag{1}$$

$$x_l = \text{MLP}(\text{LN}\,(x_l')) + x_l',\ l = 1, 2, \ldots, L \tag{2}$$

The final output can be obtained using

$$y = \text{LN}(x_L^0) \tag{3}$$

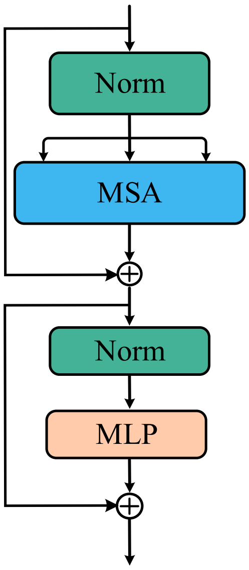

Figure 2. ViT encoder.

3.2.2. Res2Net Branch

We chose Res2Net [37] as a feature extractor to capture details and global characteristics at a finer level of granularity. In the Res2Net network, the feature maps are decomposed into four groups ($p_i, i = 1, 2, 3, 4$) from the input layer with a 1×1 convolution. The model is shown in Figure 3. The first group p_1 is not operated on, and the other groups are processed with a 3×3 convolution operation ($\kappa_i, i = 2, 3, 4$). The third and fourth groups are added with the last feature map. The above operations can be expressed as follows:

$$q_i = \begin{cases} p_i & i = 1 \\ \kappa_i(p_i) & i = 2 \\ \kappa_i(p_i + q_{i-1}) & i = 3, 4 \end{cases} \tag{4}$$

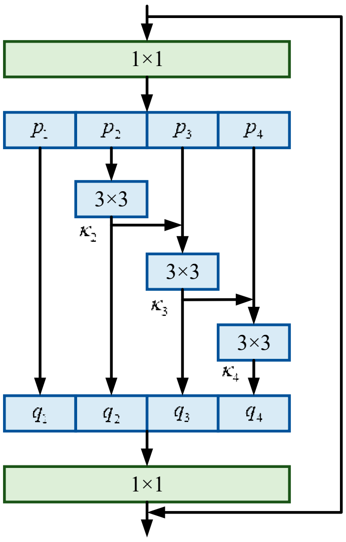

Figure 3. Res2Net network structure.

The feature maps are fused in the channel dimension. Then, the output is obtained with a 1 × 1 convolution operation. In the branch, Res2Net-50 is used to construct a feature exactor with four blocks.

3.3. Interactive Attention Mechanism

Considering the complexity of remote sensing images, a novel interactive attention mechanism is proposed to concentrate on the significant association between the two-path features and a two-branch fusion technique is utilized to maintain the integrity of the feature information.

The two features from the two branches of the feature extraction are concatenated, and they are sent to the global average pooling (GAP) layer and global max pooling (GMP) layer to produce two feature maps. GAP can focus on background information, and GMP can focus on texture information. In this block, GAP and GMP are turned into the attention weights of learning.

$$y_{\mathrm{GAP}} = \mathrm{GAP}(concat(y_1, y_2)) \tag{5}$$

$$y_{\mathrm{GMP}} = \mathrm{GMP}(concat(y_1, y_2)) \tag{6}$$

where y_1 and y_2 represent the feature maps from the ResNet and ViT branches, respectively.

The linear layer with shared weights is used to extract channel attention weights

$$y = FC(y_{\mathrm{GAP}}) \oplus FC(y_{\mathrm{GMP}}) \tag{7}$$

The weight is defined as

$$w_{\mathrm{weight}} = \mathrm{Softmax}(FC(y)) \tag{8}$$

3.4. Cross-Layer Interaction Module

In order to enhance feature utilization, a cross-layer interactive model is designed to guide the features of the two branches. This module also has two paths: the input of one path is obtained from the Res2Net branch, and the other is obtained from the ViT branch. In the first branch, the ViT-extracted features are fed into a convolution layer and concatenated with the Res2Net-extracted feature map. In the other branch, the Res2Net-extracted features are fed into a convolution layer and concatenated with the ViT-extracted feature map. The two path outputs of the cross-layer interaction module are obtained via element-wise multiplication with the weight w_{weight}, which is calculated using the interactive attention mechanism. The feature information of the two paths and the input feature are fused with the addition operation.

In order to make the network converge faster and better, the activation function uses the Gaussian error linear unit (GeLU) [38], whose expression is

$$\mathrm{GeLU}(x) = 0.5x\left(1 + \tanh\left(\sqrt{2/\pi}(x + 0.044715x^3)\right)\right) \tag{9}$$

where μ and σ are the mean and standard deviation of the normal distribution, respectively.

4. Experiments and Results

In order to verify the viability of our classification model (FCIHMRT), we performed extensive experiments and evaluations on three datasets for the categorization of general scenes. The experimental conditions and these datasets are first introduced. After that, the performance disparities between FCIHMRT and a number of cutting-edge techniques are then compared quantitatively.

4.1. Experimental Datasets

The following three widely used datasets are selected in order to test the performance of FCIHMRT:

(1) UCM [39]: This dataset is compiled by the United States Geological Survey. There are 21 classes of remote sensing scenes, involving airplanes, rivers, beaches, buildings, etc., and 100 images with a resolution of 0.3 m and a fixed size of 256 × 256 pixels make up each class. With fewer classes in the UCM data set, the distinctions between them are more pronounced. In Figure 4, a few samples from the UCM dataset are displayed.

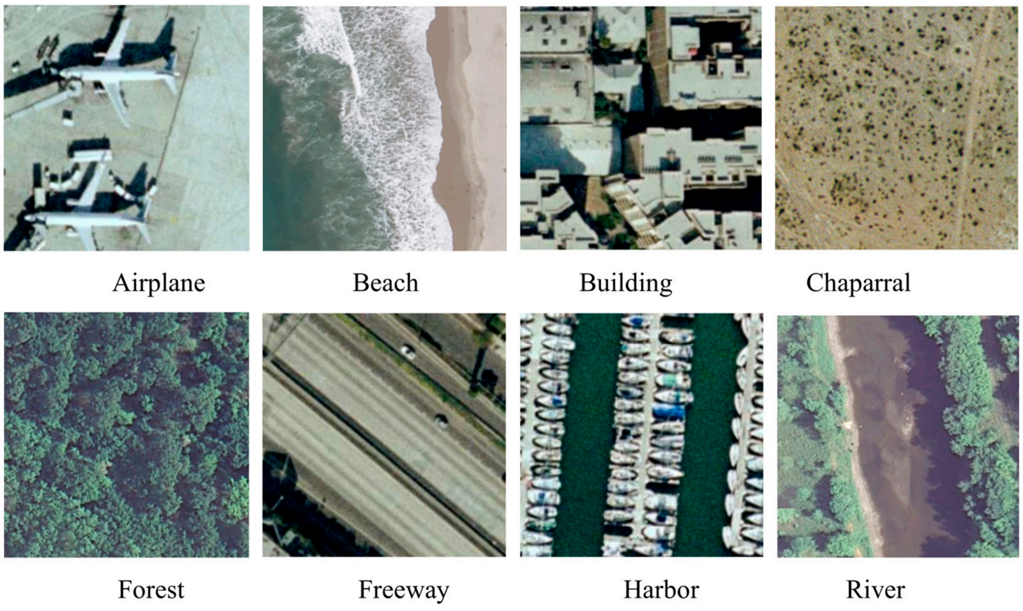

Figure 4. Some samples from the UCM dataset.

(2) AID [40]: Sample photos of this sizable dataset were compiled from Google Earth imagery. There are 30 classes in total, including resorts, bare land, and railroad stations. With between 220 and 420 images each class, there are 10,000 images total. The pixel resolution spans from 1 m to 8 m, and the image size is 600 × 600 pixels. Figure 5 displays some samples from the AID dataset.

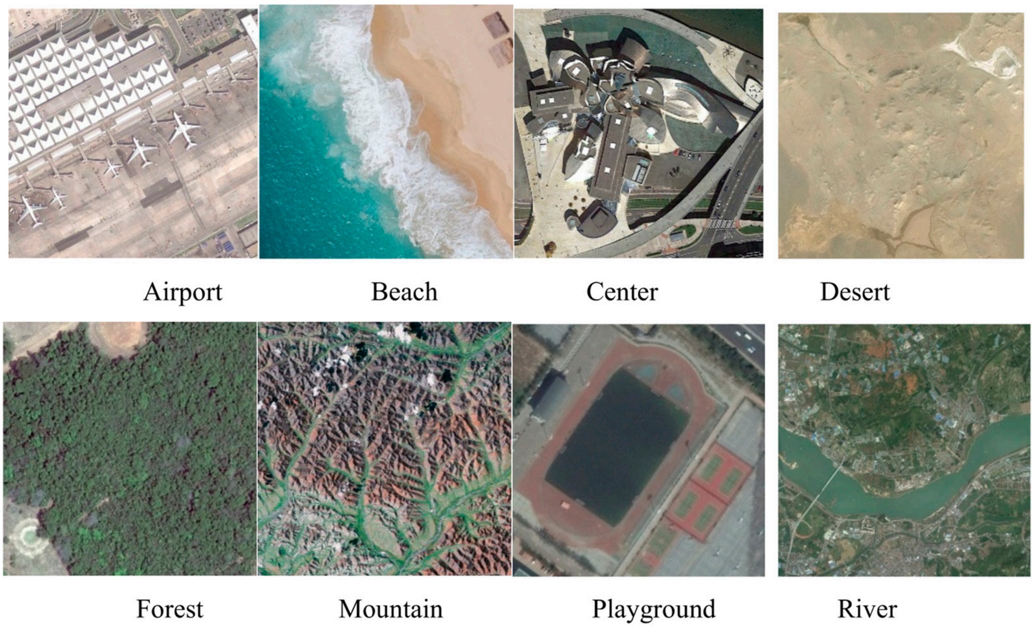

Figure 5. Some samples from the AID dataset.

(3) NWPU [41]: This is a sizable dataset that was produced by Northwestern Polytechnical University. The NWPU database contains 31,500 images with 256 × 256 pixels. The collection includes 700 images for each of the 45 classes of remote sensing scenes with a pixel resolution from 30 m to 0.2 m. Some samples from the NWPU dataset are shown in Figure 6.

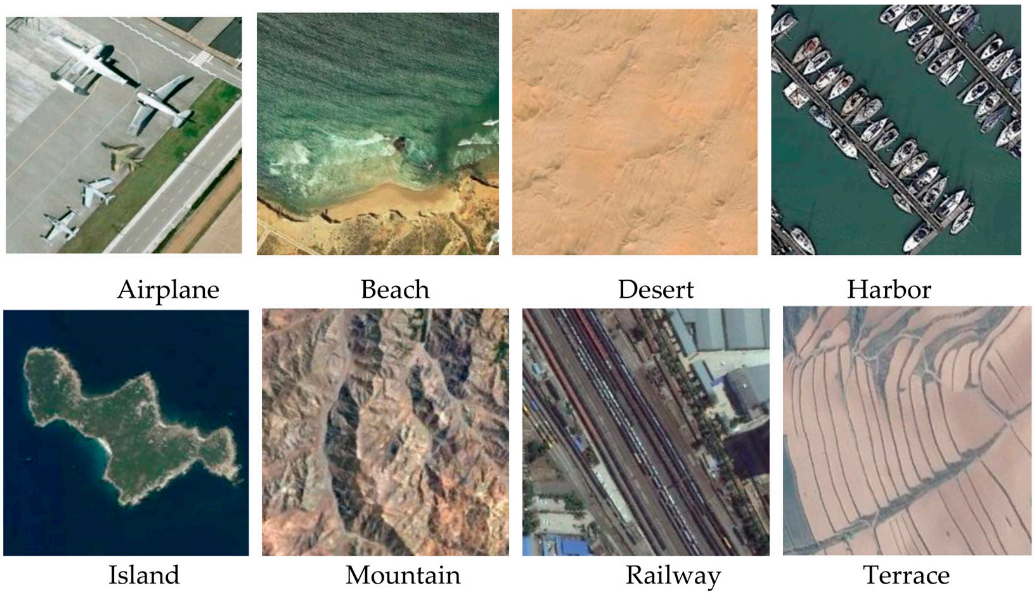

Figure 6. Some samples from the NWPU dataset.

4.2. Experimental Settings

All experiments were carried out on a computer with an Intel Core i9-10900K processor and an NVIDIA GeForce RTX 3080 GPU with 12 GB. For the UCM dataset, the training ratios were set to 80% and 50%, and the remaining portion was used for testing. The training ratios were set at 50% and 20% for the AID dataset, and the training ratios were set to 20% and 10% for the NWPU dataset. The batch size was specified as 64, and all images were resized to 224 × 224. The number of transformer layers (l) was set to 12. In order to boost the data diversity, some images were transformed during model training via random shifting, rotating, and flipping. To ensure more accurate findings, each experiment was run five times.

4.3. Evaluation Metrics

The two most often used evaluation indicators for image classification tasks are the overall accuracy (OA) and confusion matrix (CM).

OA was used to measure the performance of FCIHMRT. OA is determined as the proportion of correctly classified samples out of the total number of samples in the test set. It reflects the classification performance on the whole data set. The calculation formula is

$$OA = \frac{S}{N} \times 100\% \tag{10}$$

where S is the number of correctly classified samples in the test set and N is the total number of samples in the test set.

A CM presents the error between categories more intuitively in the form of a matrix. The value in each column of the CM denotes the number of predicted images, and the value in each row of the CM denotes the number of true images.

4.4. Experimental Results

4.4.1. Results Using UCM

FCIHMRT was compared with some of the latest remote sensing scene classification algorithms using UCM, as exhibited in Table 2. FCIHMRT outperformed the conventional scene categorization techniques when the training ratio was 80% or 50%. It can be observed that, compared with other methods, FCIHMRT had the highest OA of 99.31% and 98.84%, at training ratios of 80% and 50%, respectively. For instance, when 80% of the images were used for training, the overall accuracy of the proposed technique was roughly 0.02% greater than the overall accuracy of the ViT method without the Res2Net blocks. The OA of the proposed approach outperformed ViT by 0.09% when training with 50% of the images. In addition, FCIHMRT outperformed GAN by 0.73% and 1.30%. This demonstrates that FCIHMRT is effective in generally enhancing classification accuracy.

Table 2. Overall accuracy (%) using UCM.

Method	80% Training Ratio (OA)	50% Training Ratio (OA)
GoogLeNet [40]	94.31 ± 0.89	92.70 ± 0.60
VGG-16 [40]	95.21 ± 1.20	94.14 ± 0.69
CRAN [42]	95.75 ± 0.80	94.21 ± 0.75
MobileNet V2 [43]	99.01 ± 0.21	97.88 ± 0.31
SE-MDPMNet [44]	98.95 ± 0.12	98.36 ± 0.14
Two-Stream Fusion [45]	98.02 ± 1.03	96.97 ± 0.75
ViT [4]	99.29 ± 0.34	98.75 ± 0.21
CFDNN [46]	98.62 ± 0.27	97.65 ± 0.18
Inception-v3-CapsNet [18]	99.05 ± 0.24	97.59 ± 0.16
GSSF [47]	99.24 ± 0.47	97.86 ± 0.56
PCNet [48]	99.25 ± 0.37	98.71 ± 0.22
GAN [26]	98.58 ± 0.33	97.54 ± 0.25
Ours	99.31 ± 0.21	98.84 ± 0.26

As shown in Figure 7, the confusion matrix with all 21 classes was also created to further examine the performance of FCIHMRT with an 80% training ratio. It can be seen that the accuracy of 19 classes reached 100%. The other two classes also had an accuracy of 95%. These misclassified images may be relatively similar, and this usually reduces accuracy.

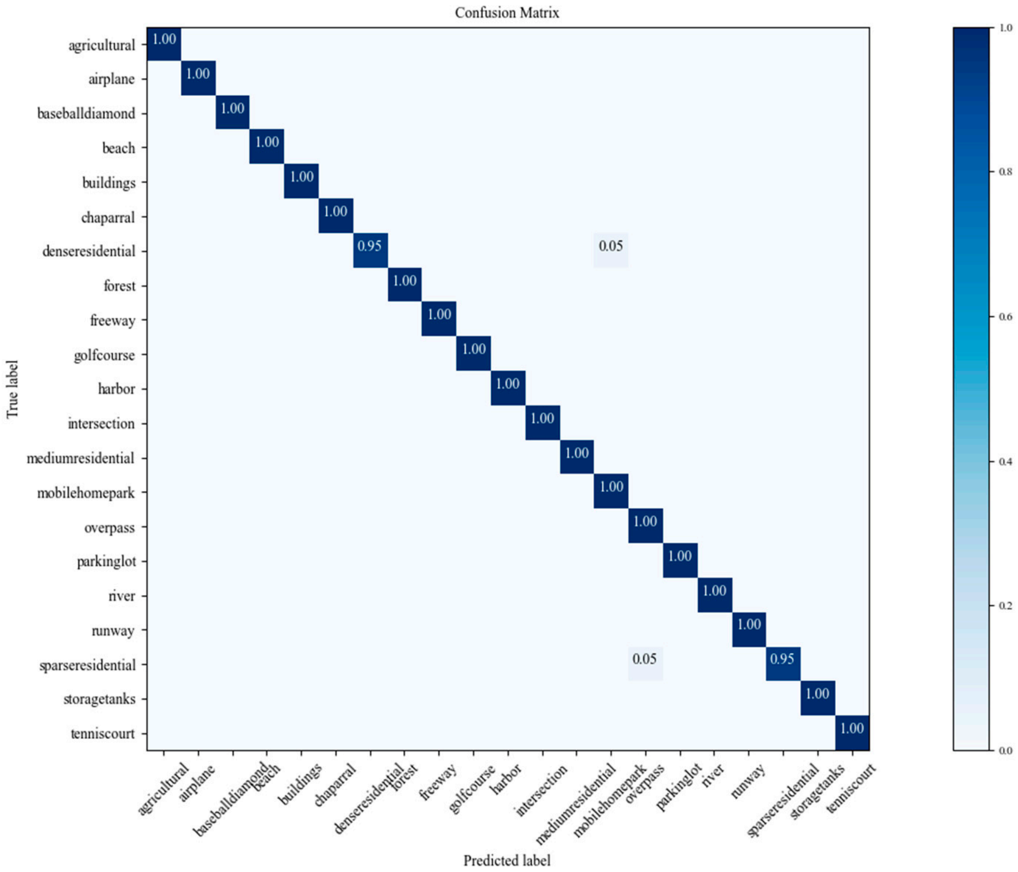

Figure 7. CM using UCM dataset and a training ratio of 80%.

4.4.2. Results Using AID

Since there are only few samples in the UCM dataset, a larger dataset must be used to evaluate FCIHMRT. Additional trials were carried out using the massive dataset AID, which consists of 10,000 images with 30 classes. The experiment was split into two parts, using 50% of the training samples or 20% of the test samples. Table 3 displays a comparison of the test results.

Table 3. Overall accuracy (%) using AID.

Method	50% Training Ratio (OA)	20% Training Ratio (OA)
GoogLeNet [40]	86.39 ± 0.55	83.44 ± 0.40
VGG-16 [40]	89.64 ± 0.36	86.59 ± 0.29
CRAN [42]	96.65 ± 0.20	95.24 ± 0.16
MobileNet V2 [43]	95.96 ± 0.27	94.13 ± 0.28
SE-MDPMNet [44]	97.14 ± 0.15	94.68 ± 0.07
Two-Stream Fusion [45]	94.58 ± 0.25	92.32 ± 0.41
ViT [4]	96.88 ± 0.19	95.58 ± 0.18
CFDNN [46]	96.56 ± 0.24	94.56 ± 0.24
Inception-v3-CapsNet [18]	96.32 ± 0.12	93.79 ± 0.13
GSSF [47]	97.65 ± 0.80	95.71 ± 0.22
PCNet [48]	96.76 ± 0.25	95.53 ± 0.16
GAN [26]	96.45 ± 0.19	94.51 ± 0.15
Ours	97.92 ± 0.29	95.82 ± 0.25

Table 3 shows that FCIHMRT produced the greatest outcomes at both 50% and 20% training rates. For instance, compared with GSSF as an outstanding method, the accuracy increased by 0.27% when 50% of the samples were used for training. When 20% of the samples were used for training, the accuracy rate increased by 0.11%. Note that the OA of FCIHMRT is much higher than that of the ViT method without the Res2Net blocks. In addition, the accuracy of FCIHMRT was 1.22% higher than that of PCNet under a training ratio of 50%. This proves that FCIHMRT is also effective with large-scale datasets.

The CM using AID with all 30 classes under a training ratio of 50% is shown in Figure 8. It can be observed from the CM that the accuracies of the industrial, park, school, and square classes were less than 95%, while the remaining classes could be accurately classified, which also proves the difficulty of this dataset. The suggested strategy could solve the problem of large intra-class differences and small inter-class differences of remote sensing images to a certain extent.

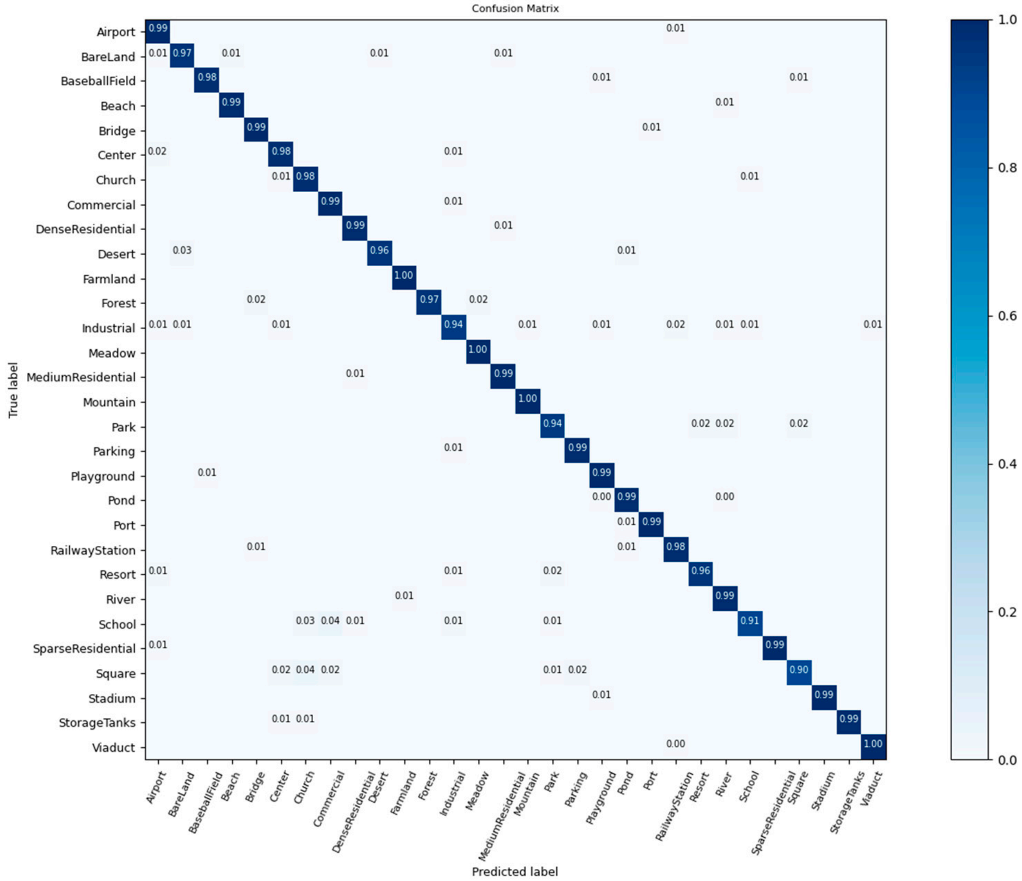

Figure 8. CM using AID dataset and training ratio of 50%.

4.4.3. Results Using NWPU

In order to validate the generalization ability of FCIHMRT, NWPU with 45 classes can be used as a multiclass database. Table 4 compares the classification accuracy of FCIHMRT with that of other methods using the NWPU dataset. Only a 93.63% accuracy was obtained using GAN. When the training ratio was 20%, the accuracy was 94.86%, which is higher than that of the ViT and PCNet models. This further verifies that FCIHMRT can effectively obtain the deep features of scenes with a stronger expression ability.

Figure 9 reveals that the classification accuracies of all 45 classes in NWPU obtained using FCIHMRT with a 20% training ratio were higher than 80%. Among them, the classification accuracy of the golf course and mobile home park classes reached 99%, which indicates that FCIHMRT has a good classification performance for scenes with a small feature complexity. Meanwhile, it was demonstrated that the OA of palace scenes was reduced to 80%, in which some palace scenes were categorized into the church intersection and island classes, indicating that the classification capacity of FCIHMRT still requires improvement for similar scenes but in general, it can distinguish different scenes with rich spatial information.

Table 4. Overall accuracy (%) using NWPU.

Method	20% Training Ratio (OA)	10% Training Ratio (OA)
GoogleNet [40]	78.48 ± 0.26	76.19 ± 0.38
VGG-16 [40]	79.79 ± 0.15	76.47 ± 0.18
CRAN [42]	94.07 ± 0.08	91.28 ± 0.19
MobileNet V2 [43]	83.26 ± 0.17	80.32 ± 0.16
SE-MDPMNet [44]	94.11 ± 0.03	91.80 ± 0.07
Two-Stream Fusion [45]	83.16 ± 0.18	80.22 ± 0.22
ViT [4]	94.50 ± 0.18	91.17 ± 0.13
CFDNN [46]	93.83 ± 0.09	91.17 ± 0.13
Inception-v3-CapsNet [18]	92.6 ± 0.11	89.03 ± 0.21
GSSF [47]	94.48 ± 0.26	91.98 ± 0.19
PCNet [48]	94.59 ± 0.07	92.64 ± 0.13
GAN [26]	93.63 ± 0.12	91.06 ± 0.11
Ours	94.86 ± 0.21	92.67 ± 0.26

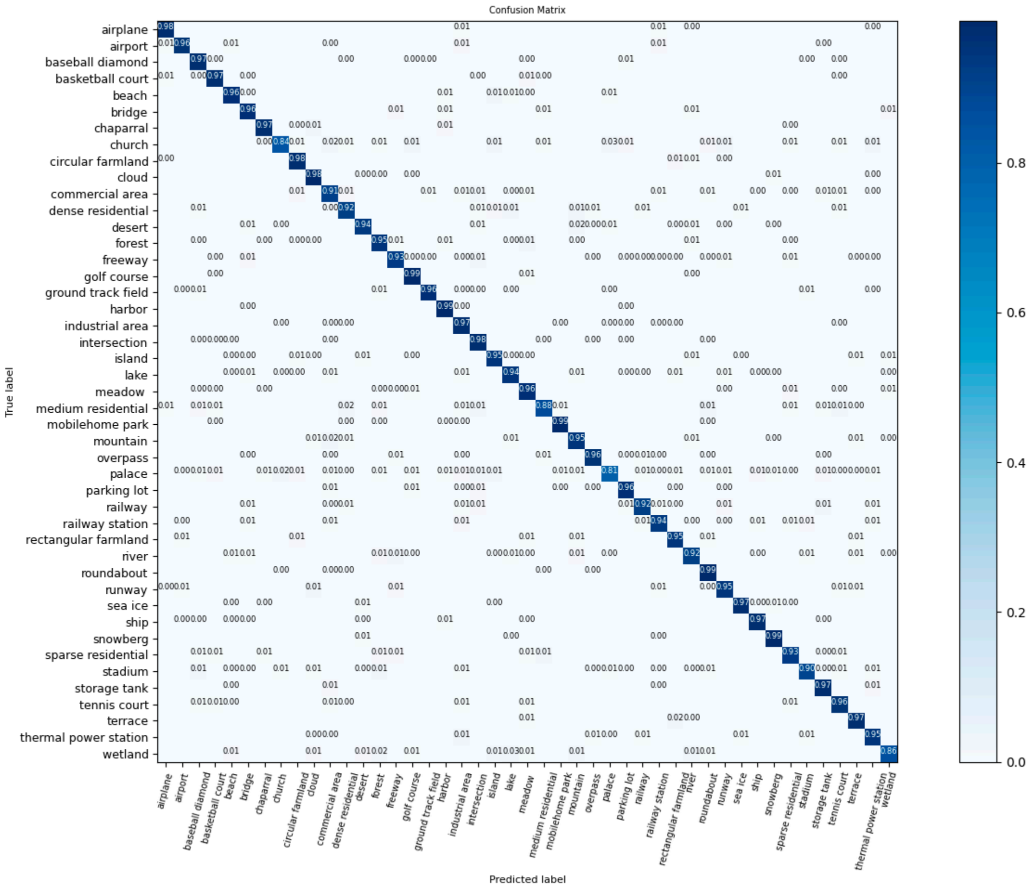

Figure 9. CM using NWPU dataset and a training ratio of 20%.

5. Conclusions

In this paper, a new network model (FCIHMRT) was proposed for remote sensing scene classification. In the overall network model, the features of Res2Net and ViT are obtained by using a two-channel structure as the feature extractor. The feature information extracted from the two branches is fused to enlarge the receptive field and enrich the scene information of remote sensing features, effectively overcoming the shortcoming of CNNs using a single fixed-size convolution kernel. The proposed interactive attention mechanism enhances the model's focus on the key regions of the image and avoids the interference caused by redundant background details. By using the cross-layer fusion module to fuse multilevel features, a distinguishable and robust fusion feature representation can be obtained. Finally, a performance improvement over the current approaches was demonstrated by training and testing them using three commonly used datasets: UCM, AID, and NWPU. The results of the experiment demonstrate that FCIHMRT is better suited to the task of scene classification. However, it should be noted that FCIHMRT can achieve a high classification accuracy, but it is still limited by increased computation time. In future work, from the perspective of accelerating network training, we will focus on constructing efficient and high-precision models, which can be combined with lightweight networks and data dimension reduction methods.

Author Contributions: Conceptualization, Y.H. and S.G.; methodology, Y.H. and C.G.; software, Y.H. and S.G.; validation, S.G.; investigation, Y.H.; writing—original draft preparation, Y.H. and S.G.; writing—review and editing, C.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (grant number 42101017), the China Postdoctoral Science Foundation (grant number 2020M680979), and the funding project of the Northeast Geological S&T Innovation Center of China Geological Survey (grant numbers QCJJ2022-23, QCJJ2022-24, and QCJJ2022-26).

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

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

References

1. Ghaffarian, S.; Valente, J.; van der Voort, M.; Tekinerdogan, B. Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review. *Remote Sens.* **2021**, *13*, 2965. [CrossRef]
2. Cheng, G.; Xie, X.; Han, J.; Guo, L.; Xia, G.-S. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 3735–3756. [CrossRef]
3. Xu, C.; Zhu, G.; Shu, J. A combination of lie group machine learning and deep learning for remote sensing scene classification using multi-layer heterogeneous feature extraction and fusion. *Remote Sens.* **2022**, *14*, 1445. [CrossRef]
4. Xu, K.; Deng, P.; Huang, H. Vision transformer: An excellent teacher for guiding small networks in remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618715. [CrossRef]
5. Zhu, Q.; Zhong, Y.; Zhao, B.; Xia, G.-S.; Zhang, L. Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* **2016**, *13*, 747–751. [CrossRef]
6. 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]
7. Chen, L.; Li, S.; Bai, Q.; Yang, J.; Jiang, S.; Miao, Y. Review of image classification algorithms based on convolutional neural networks. *Remote Sens.* **2021**, *13*, 4712. [CrossRef]
8. Ao, L.; Feng, K.; Sheng, K.; Zhao, H.; He, X.; Chen, Z. Tpenas: A two-phase evolutionary neural architecture search for remote sensing image classification. *Remote Sens.* **2023**, *15*, 2212. [CrossRef]
9. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv* **2014**, arXiv:1409.0473.
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017.
11. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S. An image is worth 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
12. Deng, P.; Xu, K.; Huang, H. When cnns meet vision transformer: A joint framework for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8020305. [CrossRef]
13. Gu, J.; Wang, Z.; Kuen, J.; Ma, L.; Shahroudy, A.; Shuai, B.; Liu, T.; Wang, X.; Wang, G.; Cai, J. Recent advances in convolutional neural networks. *Pattern Recognit.* **2018**, *77*, 354–377. [CrossRef]
14. Kaul, A.; Kumari, M. A literature review on remote sensing scene categorization based on convolutional neural networks. *Int. J. Remote Sens.* **2023**, *44*, 2611–2642. [CrossRef]
15. Pires de Lima, R.; Marfurt, K. Convolutional neural network for remote-sensing scene classification: Transfer learning analysis. *Remote Sens.* **2020**, *12*, 86. [CrossRef]
16. 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]

17. Li, E.; Samat, A.; Du, P.; Liu, W.; Hu, J. Improved bilinear cnn model for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 8004305. [CrossRef]
18. Zhang, W.; Tang, P.; Zhao, L. Remote sensing image scene classification using cnn-capsnet. *Remote Sens.* **2019**, *11*, 494. [CrossRef]
19. Peng, F.; Lu, W.; Tan, W.; Qi, K.; Zhang, X.; Zhu, Q. Multi-output network combining gnn and cnn for remote sensing scene classification. *Remote Sens.* **2022**, *14*, 1478. [CrossRef]
20. Huang, X.; Zhou, Y.; Yang, X.; Zhu, X.; Wang, K. Ss-tmnet: Spatial–spectral transformer network with multi-scale convolution for hyperspectral image classification. *Remote Sens.* **2023**, *15*, 1206. [CrossRef]
21. Wang, J.; Luo, C.; Huang, H.; Zhao, H.; Wang, S. Transferring pre-trained deep cnns for remote scene classification with general features learned from linear pca network. *Remote Sens.* **2017**, *9*, 225. [CrossRef]
22. Zhao, H.; Liu, F.; Zhang, H.; Liang, Z. Convolutional neural network based heterogeneous transfer learning for remote-sensing scene classification. *Int. J. Remote Sens.* **2019**, *40*, 8506–8527. [CrossRef]
23. Wang, W.; Chen, Y.; Ghamisi, P. Transferring cnn with adaptive learning for remote sensing scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5533918. [CrossRef]
24. Xu, S.; Mu, X.; Chai, D.; Zhang, X. Remote sensing image scene classification based on generative adversarial networks. *Remote Sens. Lett.* **2018**, *9*, 617–626. [CrossRef]
25. Han, W.; Wang, L.; Feng, R.; Gao, L.; Chen, X.; Deng, Z.; Chen, J.; Liu, P. Sample generation based on a supervised wasserstein generative adversarial network for high-resolution remote-sensing scene classification. *Inf. Sci.* **2020**, *539*, 177–194. [CrossRef]
26. Ma, A.; Yu, N.; Zheng, Z.; Zhong, Y.; Zhang, L. A supervised progressive growing generative adversarial network for remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5618818. [CrossRef]
27. Zhao, Z.; Li, J.; Luo, Z.; Li, J.; Chen, C. Remote sensing image scene classification based on an enhanced attention module. *IEEE Geosci. Remote Sens. Lett.* **2021**, *18*, 1926–1930. [CrossRef]
28. 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]
29. Wang, D.; Lan, J. A deformable convolutional neural network with spatial-channel attention for remote sensing scene classification. *Remote Sens.* **2021**, *13*, 5076. [CrossRef]
30. Tian, T.; Li, L.; Chen, W.; Zhou, H. Semsdnet: A multiscale dense network with attention for remote sensing scene classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 5501–5514. [CrossRef]
31. Wang, D.; Zhang, C.; Han, M. Mlfc-net: A multi-level feature combination attention model for remote sensing scene classification. *Comput. Geosci.* **2022**, *160*, 105042. [CrossRef]
32. Shen, J.; Yu, T.; Yang, H.; Wang, R.; Wang, Q. An attention cascade global–local network for remote sensing scene classification. *Remote Sens.* **2022**, *14*, 2042. [CrossRef]
33. Yu, Y.; Li, Y.; Wang, J.; Guan, H.; Li, F.; Xiao, S.; Tang, E.; Ding, X. C^2-capsvit: Cross-context and cross-scale capsule vision transformers for remote sensing image scene classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6512005. [CrossRef]
34. Zhang, J.; Zhao, H.; Li, J. Trs: Transformers for remote sensing scene classification. *Remote Sens.* **2021**, *13*, 4143. [CrossRef]
35. Sha, Z.; Li, J. Mitformer: A multiinstance vision transformer for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 6510305. [CrossRef]
36. Wang, G.; Chen, H.; Chen, L.; Zhuang, Y.; Zhang, S.; Zhang, T.; Dong, H.; Gao, P. P 2fevit: Plug-and-play cnn feature embedded hybrid vision transformer for remote sensing image classification. *Remote Sens.* **2023**, *15*, 1773. [CrossRef]
37. Gao, S.H.; Cheng, M.M.; Zhao, K.; Zhang, X.Y.; Yang, M.H.; Torr, P. Res2net: A new multi-scale backbone architecture. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 652–662. [CrossRef]
38. Hendrycks, D.; Gimpel, K. Gaussian error linear units (gelus). *arXiv* **2016**, arXiv:1606.08415.
39. 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.
40. 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]
41. 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]
42. Wang, Y.; Hu, Y.; Xu, Y.; Jiao, P.; Zhang, X.; Cui, H. Context residual attention network for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 8022805. [CrossRef]
43. Pan, H.; Pang, Z.; Wang, Y.; Wang, Y.; Chen, L. A new image recognition and classification method combining transfer learning algorithm and mobilenet model for welding defects. *IEEE Access* **2020**, *8*, 119951–119960. [CrossRef]
44. Zhang, B.; Zhang, Y.; Wang, S. A lightweight and discriminative model for remote sensing scene classification with multidilation pooling module. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 2636–2653. [CrossRef]
45. Yu, Y.; Liu, F. A two-stream deep fusion framework for high-resolution aerial scene classification. *Comput. Intell. Neurosci.* **2018**, *2018*, 8639367. [CrossRef] [PubMed]
46. Deng, P.; Huang, H.; Xu, K. A deep neural network combined with context features for remote sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2020**, *19*, 8000405. [CrossRef]

47. Gao, Y.; Sun, X.; Liu, C. A general self-supervised framework for remote sensing image classification. *Remote Sens.* **2022**, *14*, 4824. [CrossRef]
48. Zhang, Y.; Zheng, X.; Lu, X. Pairwise comparison network for remote-sensing scene classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 6505105. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.